ENHANCED DNA ENCODING ALGORITHM FOR ANOMALY INTRUSION DETECTION SYSTEM

OMAR FITIAN RASHID AL-RAWI

THESIS SUBMITTED IN FULFILMENT FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

FACULTY OF INFORMATION SCIENCE AND TECHNOLOGY
UNIVERSITI KEBANGSAAN MALAYSIA
BANGI

2019

PENINGKATAN ALGORITMA PENGEKODAN DNA UNTUK SISTEM PENGESANAN PENCEROBOHAN ANOMALI

OMAR FITIAN RASHID AL-RAWI

TESIS YANG DIKEMUKAKAN UNTUK MEMPEROLEHI IJAZAH
DOKTOR FALSAFAH

FAKULTI TEKNOLOGI DAN SAINS MAKLUMAT
UNIVERSITI KEBANGSAAN MALAYSIA
BANGI

2019

DECLARATION

I hereby declare that the work in this thesis is my own except for quotations and summaries which have been duly acknowledged.

24 April 2019

OMAR FITIAN RASHID
P80374

ACKNOWLEDGEMENTS

First and foremost, praise be to Almighty Allah for all his blessings for giving me patience and good health throughout the duration of this PhD research.

I would like to acknowledge here the unflagging efforts of my supervisors, Asst. Prof. Dr. Zulaiha Ali Othman and my co-supervisor, Dr. Suhaila Zainudin, I would like to express my deep gratitude for their supervision, advice, and encouragement throughout my research.

My thanks and respect to my father Dr. Fitian Rashid, my mother Dr. Khawla Hori, and my wife Ola Oqbah for their supports during the period of my study.

ABSTRACT

An intrusion detection system (IDS) aims to identify unauthorized use, misuse, and abuse of computer systems or network. Usually, the quality of IDS is measured based on detection rate (DR) and false alarm rate (FAR). Machine learning is the most popular technique used for intrusion detection system. Various good algorithms are proposed, obtaining high detection rate by improving the algorithm or hybridizing with other algorithm, however; they are still suffering with the time especially after the improvement of the algorithm and dealing with large traffic data. On other hand, previous researches have successfully applied the Deoxyribonucleic Acid (DNA) approaches for misuse and anomaly intrusion detection system. However, the results showed very low detection rate with low processing time. Literature review have found that three factors influence the quality solution of DNA disease detection: DNA encoding method, DNA's Keys and their positions and matching method used to detect anomaly. Therefore, the aim of this research is to propose a suitable DNA approach for anomaly IDS with two objectives. The first objective; is to propose an enhanced three DNA encoding algorithms for anomaly intrusion detection system using two keys and their positions (DNA-IDS) known as DEM4all, DEM3sel, and DEMdif. These encoding methods are used to convert the network traffic dataset into a form of DNA sequences. DEM4all used the same characters number to represent all attributes, DEM3sel used three characters to represent all attributes and put a single character to distinguish between nominal and numerical attributes; while DEMdif used different number of characters to represent all attributes based on attributes values and put a single character to distinguish between nominal and numerical attributes. The second objective; is to improve the best propose DNA encoding approach, using three and four DNA's keys and their positions, applied four matching algorithms and also combination of two algorithms, and propose new features selection method based on DNA location. The experiments are conducted using the KDDCup'99 and NSL-KDD datasets. The results showed that DEM3sel encoding method is obtained the best results, with detection rate for DoS, Probe, R2L, and U2R as 99.54%, 99.87%, 99.99%, and 100% respectively, with DR, FAR, and accuracy up to 99.58%, 35.53%, and 92.74% respectively, and matching time equal to 62 seconds. Then an improvement method, used three and four keys and their positions, has improved the accuracy up to 1.75%, and reduced FAR up to 26.48% and matching time to 74 seconds. Furthermore, the implementation of four matching algorithms (Brute force algorithm, Boyer Moore algorithm, Horspool algorithm and Knuth-Morris-Pratt algorithm) and the combination of two algorithms (Boyer Moore Algorithm and Knuth–Morris–Pratt Algorithm) have improved the matching times up to 13,8,6,5, and 16 seconds respectively. The Knuth-Morris-Pratt algorithm has shown the best matching algorithm. The proposed features selection method has improved the DR, FAR and accuracy results up to 1.55%, 0.13% and 1.27% respectively, while the encoding time reduced from 46702 seconds using all 41 features to 325 seconds using the selected features. With such result, it can be concluded that the proposed DNA approach for IDS can be used as a good approach for effective and efficient anomaly IDS.

ABSTRAK

Sistem pengesanan pencerobohan (SPP) adalah bertujuan untuk mengenal pasti capaian tanpa kebenaran, penyalahgunaan, dan serangan ke atas sistem atau rangkaian komputer. Kebiasaannya, kualiti SPP diukur berdasarkan kadar pengesanan dan kadar penggera palsu. Pembelajaran mesin adalah merupakan teknik paling popular yang digunakan untuk sistem pengesanan pencerobohan. Pelbagai algoritma bagus telah dicadangkan, memperoleh kadar pengesanan yang tinggi dengan menambah baik algoritma atau hibridisasi dengan algoritma lain. Walau bagaimanapun, mereka menghadapi masalah dengan masa pemprosesan, terutama sekali setelah algoritma ditambah baik dan berurusan dengan data lalu lintas yang besar. Selain itu, penyelidikan lepas juga telah berjaya mengaplikasi pendekatan DNA untuk sistem pengesanan penyalahgunaan dan pencerobohan anomali. Walau bagaimanapun, keputusannya mengalami kadar pengesanan yang sangat rendah dengan masa pemprosesan yang cepat. Kajian kesusasteraan telah mendapati bahawa terdapat tiga faktor yang mempengaruhi penyelesaian kualiti pengesanan penyakit DNA: kaedah pengekodan DNA, kekunci dan kedudukan DNA serta kaedah padanan yang digunakan untuk pengesanan anomali. Oleh itu, kajian ini bertujuan untuk mencadangkan peningkatan algoritma pengekodan DNA yang sesuai untuk SPP anomali dengan tiga objektif utama. Pertama, mencadangkan tiga algoritma pengekodan DNA baru untuk sistem pengesanan pencerobohan anomali menggunakan dua kekunci dan kedudukan DNA (DNA-IDS) mereka yang dikenali sebagai DEM4all, DEM3sel, dan DEMdif. Kaedah pengekodan ini digunakan untuk menukar kumpulan data rangkaian trafik ke dalam bentuk urutan DNA. DEM4all menggunakan empat aksara yang untuk mewakili semua atribut, DEM3sel menggunakan tiga aksara untuk mewakili semua atribut dan meletakkan satu aksara untuk membezakan antara atribut nominal dan berangka, manakala DEMdif menggunakan aksara yang berbeza untuk mewakili kesemua atribut berdasarkan nilai atribut dan meletakkan aksara tunggal untuk membezakan antara atribut nominal dan berangka. Kedua, memperbaiki pendekatan DNA cadangan terbaik sebelumnya menggunakan kekunci DNA tiga dan empat serta kedudukan DNA, dan menggunakan empat algoritma padanan dan juga gabungan dua algoritma. Ketiga, menambah baik algoritma menggunakan kaedah pemilihan ciri berdasarkan lokasi DNA. Eksperimen dijalankan menggunakan set data KDDCup'99 dan NSL-KDD. Keputusan menunjukkan bahawa kaedah pengekodan DEM3sel memperoleh keputusan terbaik, dengan kadar pengesanan untuk DoS, Probe, R2L, dan U2R masing-masing sebanyak 99.54%, 99.87%, 99.99% dan 100%, dengan DR, FAR dan ketepatan masing-masing meningkat sehingga 99.58%, 35.53%, dan 92.74%, dalam tempoh masa 62 saat yang hampir sama. Penambahbaikan algoritma menggunakan tiga dan empat kekunci dan kedudukan DNA telah meningkatkan ketepatan sehingga 1.75%, dan menurunkan FAR sehingga 26.48%, dalam masa 74 saat yang hampir sama. Tambahan pula, empat algoritma padanan digunakan iaitualgoritma kekerasan, algoritma Boyer Moore, algoritma Horspool dan algoritma Knuth-Morris-Pratt. Gabungan dua algoritma (algoritma Boyer Moore dan algoritma Knuth-Morris-Pratt) telah menambah baik masa sehingga 13,8,6,5 dan 16 saat, namun algoritma Knuth-Morris-Pratt telah menunjukkan sebagai algoritma padanan terbaik. Penggunaan pemilihan ciri telah meningkatkan keputusan DR, FAR dan ketepatan masing-masing sehingga 1.55%, 0.13% dan 1.27%, manakala masa pemprosesan dikurangkan daripada 46702 saat kepada 325 saat menggunakan semua 41 ciri yang dicadangkan. Hasilnya, dapat disimpulkan bahawa pendekatan DNA yang dicadangkan untuk IDS boleh digunakan sebagai satu pendekatan yang baik untuk PPS anomali yang berkualiti dan pantas.

CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

LIST OF ABBREVIATIONS

DNA	Deoxyribonucleic Acid
DoS	Denial of Service
DR	Detection Rate
EP	Elementary Patterns
FAR	False Alarm Rate
FN	False Negative
FP	False Positive
HIDS	Host Intrusion Detection System
IDS	Intrusion Detection System
KDDCup 99	Knowledge Discovery and Data mining
NIDS	Network Intrusion Detection System
NSL-KDD	Network Security Laboratory-Knowledge Discovery and Data Mining
PCA	Principal Component Analysis
PCR	Polymerase Chain Reaction
R2L	Remote to Local
RFLP	Restriction Fragment Length Polymorphism
STR	Short Tandem Repeats
TN	True Negative
TP	True Positive
U2R	User to Root

CHAPTER I

INTRODUCTION

1.1 INTRODUCTION

Security plays an important role in the usage of computer network by people and companies is getting more widespread (Adlakha & Subramanium 2013). The growth of using internet involves the availability of information. On the other hand, the new attackers use the internet to achieve their goals (Moore et al. 2003). Several types of computer security like antivirus and firewall are now routinely used. Antivirus is a computer software used to prevent, detect and remove malicious software (Naveen 2016). These can be applied to protect the network users from malwares (viruses, Trojan horses, worms and spywares) known with a signature stored in their database. Signature files of many anti-virus products are updated only on a weekly or daily basis. Therefore, computer users are exposed to new intrusion during the intervals between updates. The firewall is a system that monitors and controls the incoming and outgoing network traffic based on predetermined security rules (Boudriga 2010). The flaw of firewall is that it can only prevent external intrusion. This motivate the companies to build set of their own monitoring system that monitors data flow in their network. This system is called Intrusion Detection Systems (IDS).

The intrusion detection systems are used to identify unauthorized user and misuse of computer systems and network. IDS can identify both outside and inside attackers (Adlakha & Subramanium 2013). An IDS deals with huge amount of data which contains irrelevant and redundant features causing slow training and testing process, higher resource consumption as well as poor detection rate. Feature selection is one of the key topics in IDS.

An IDS needs to check a huge number of data; therefore, IDS has many challenges such as low detection rate, complicated system and time. Therefore, in order to deal with these challenges, a new DNA encoding method based on DNA technique with feature selection method is proposed which helps to build a lightweight IDS with high detection rate and improved detection time.

The performance of intrusion detection system is measured based on four factors. These are, the detection rate (DR), which is the ratio of the number of correctly detected attacks to the total number of attacks, the false alarm rate (FAR), is the ratio of the number of normal connections that are incorrectly misclassified as attacks to the total number of normal connections. The accuracy is measuring the ratio of the number of truly classified connections to the total number of connections (Wu & Benzhaf 2010). Finally, the execution time includes both encoding time and matching time.

The purpose of this chapter is to discuss the background of solving the IDS problems presented in Section 1.2. The problem statement is shown in Section 1.3. The research questions are in Section 1.4. The research objectives are presented in Section 1.5. The significance of the research is described in Section 1.6. The scope of the research is described in Section 1.7. The research methodology is summarized in Section 1.8. Finally, Section 1.9 outline the organization of the whole thesis chapters.

1.2 RESEARCH BACKGROUND

An IDS has become a primary component in computer or network security. Prevention of intrusions is entirely depending on detection capability of these systems. Regular method on IDS use Artificial Neural Networks (Hu et al. 2015; Ingre & Yadav 2015; Tammi et al. 2015), which is an information processing model that is inspired by the biological nervous systems. However, the main drawbacks of applying Artificial Neural Network for IDS are not suitable for real time because the training time is slow and the whole system must be retrained when added new attack (Shah & Trivedi 2012). Another IDS is using Artificial Immune Systems (Hosseinpour et al. 2014; Kumar & Reddy 2014; Abas et al. 2015) which is designed for the

computational system and inspired by the human immune system, but the main drawbacks of these system are complexity and high FAR results (Shen 2012). Some IDS is using Evolutionary Techniques (Thamilarasu 2015; Qiang et al. 2016; Varma et al. 2016) which is a subfield of artificial intelligence inspired from natural evolution, but the main limitation is that incorrect threshold value might lead to a high FAR result and time-consuming (Tewatia & Mishra 2015). Most IDS is using machine learning techniques (Sreenath & Udhayan 2015; Jabbar et al. 2017; Mkuzangwe & Nelwamondo 2017), but the main drawback is the high consumption of resources (Kaur et al. 2013). Finally, some IDS is using Pattern Recognition (Onik et al. 2015; Ravale et al. 2015; Li et al. 2016) to provide a reasonable answer for all possible inputs and to carry out matching of the inputs, taking into consideration their statistical variations.

IDS can be classified as anomaly detection and misuse detection. In anomaly detection, the system builds a profile that can be regarded as normal or expected usage patterns over a period of time and triggers alarms for anything that deviates from this behavior. On the other hand, in misuse detection, the system identifies the intrusions based on known intrusion techniques and triggers alarms by detecting known exploits or attacks based on their attack signatures (Depren et al. 2005). The anomaly IDS is better than misuse IDS; particularly as this system can detect new attacks (Jose et al. 2018).

Mahdy & Saeb 2007; Al-Ibaisi et al. 2008; Hameed and Rashid 2014 have proposed IDS based on DNA. The DNA is the genetic material found in most organisms (include human), since the information in DNA is stored as a code made of four chemical bases. These bases called Adenine, Cytosine, Guanine, and Thymine, and referred as A, C, G, and T (Soram & Khomdram 2010). There are several methods for IDS based on DNA, IDS is used to detect attack when it is trying to enter to the system, by looking for suspicious patterns (that different from user behavior). In DNA, the human body is looking at specific DNA sequences to check if it changes (different from the DNA of a healthy body) which means a disease is attacking the human body. Mutations detection can be achieved based on identified genetic analysis results using DNA markers such as short tandem repeat (STR) (Nakamura 2009).

An anomaly IDS is introduced by Mahdy and Saeb (2007) through using the concept of a DNA sequence or gene, which is responsible for the normal network traffic patterns. This system is achieved based on two steps, first step is network traffic DNA sequence generation, or the offline training phase and the second step is network monitoring or the online monitoring phase. Al-Ibaisi et al. (2008) generated a normal signature sequence and alignment threshold value from processing the system training data and encoded observed network connection into corresponding DNA nucleotides sequence. This system is carried out by applied two steps, first step is included the DNA sequence encoding and in the second is the genetic algorithm that used to optimize the selection for target solution. Hameed and Rashid (2014) introduced a misuse IDS based on DNA sequence. Three steps are followed to perform this system. In the first step, the network traffic is converted to DNA sequence; and in the second step, the attack signature keys and their positions are extracted by using Teiresias algorithm. In the last step, the network traffic is classified either attack or normal based on attack signature key and their positions by using Horspool algorithm. The result shows that the performance of DNA approach was fast, but the DR and accuracy results were low in comparative with machine learning approach.

There is an emerge need for IDS to be lightweight with high DR. Therefore, feature selection methods are used to reduce the computation and model complexity. These methods are divided into three categories named filter, wrapper and hybrid method (Amrita & Ahmed 2012).

1.3 PROBLEM STATEMENT

Previous literatures on applying DNA sequence in IDS have found three factors influence the IDS detection rate; which are encoding method (Al-Ibaisi et al. 2008), the keys and positions of the DNA sequence (Hameed & Rashid 2014), and matching methods (Dagar et al. 2016).

According to the first factor, Al-Ibaisi et al. (2008) proposed a DNA encoding method for each type of network attributes in a form of tables by divided the network traffic attributes into static parameters and dynamic parameters, then used three DNA characters for representation and put three characters as a header in front of static parameters attributes values. However, the detection rate results were very low, they have obtained detection rate of Denial of Service (DoS) attacks, Probe attack, Remote to Local (R2L) attacks and User to Root (U2R) attacks as 51.83%, 57.28%, 24.20%, and 43.10% respectively (Al-Ibaisi et al. 2008), where their DNA encoding has two drawbacks; the first one was the representation of Boolean (0 or 1) and integer (0-9) attributes values based on different DNA sequences (but both attributes contain same values 0 and 1). However, they put a header to distinguish between nominal and numerical attribute. In the second one, they used three characters as groups' identifiers and they tried to useless space during encoding, but only one character can be used instead of three characters. Later, various encoding methods have been proposed, Wang & Zhang (2009) have proposed a mixed DNA encoding method, that used three DNA characters to represent each alphabetic character. Then, they used two numbers to represent each DNA character (that mean used different characters number for encoding) for cryptography; where cryptography is used to secure communication of data across the open internetwork by encoding and decoding the plaintext (UbaidurRahman et al. 2015). On the other hand, Jarold et al. (2013) have used three DNA characters to represent each value for message cryptography. Hameed & Rashid (2014) enhanced the DNA encoding method that proposed by Al-Ibaisi et al. (2008) and applied it for misuse IDS. They used three DNA characters to represent the discrete attributes that contain nominal values with three characters as a header and also, they used two characters to represent the digit attributes values, the main drawback of their DNA encoding is they used three characters as group identifiers but only one character can be used instead of three characters. While UbaidurRahman et al. (2015) have used 4 DNA characters to represent each value as text cryptography, the results revealed that, the less complex DNA leads the more accuracy and detection rate in IDS. However too simple DNA sequences might not represent the best accuracy as well, due to the loss of important information. Therefore, it is a challenge to have a best DNA sequence for IDS.

In relation to the second factor, Nakamura (2009), has done many research in mutations detection in DNA that has STR which include both keys and their positions has significant important on mutations detection. This is because STR is special areas of DNA but they can be varied in different people, and when a mutation is happened it leads to permanent change in the DNA sequence and makes this sequence differs from the sequence that exist in most people. Later, Hameed & Rashid (2014) have used Teiresias algorithm to discover the STR sequence for network traffic based on attack signature. The achieved results showed that the use of keys only or keys and their positions have an important effect on the performance of the proposed method. Therefore, in order to have better detection rate in IDS, it needs to identify the STR and their positions, their method have some limitations such as they used 10% KDD dataset for both training and testing (no possibility to check new attacks), used 4000 random records only for testing (the performance evaluations is not accurate compare with using all records), used eight keys to distinguished between normal records and attacks records (required high matching time), and used misuse to detect attack while disease detection in human body follow the anomaly detection idea.

The third factor is concerned with matching algorithms. Al-Ibaisi et al. (2008) have used genetic algorithm to find DNA signature but the achieved results are low. The results achieved when applied matching algorithm on Al-Ibaisi et al. (2008) are higher than using genetic algorithm that used in their method. On another hand, matching algorithms are applied to detect attacks in intrusion detection system, various pattern matching algorithms are used for intrusion detection system (Raj et al. 2015; Ravale et al. 2015; Sonawane & Pattewar 2015; Chen et al. 2016; Farnaaz & Jabbar 2016; Promod & Jacob 2016). Various algorithms are applied for intrusion detection system; some of these algorithms are Brute Force Algorithm, Boyer Moore Algorithm, Horspool Algorithm, and Knuth–Morris–Pratt Algorithm. These algorithms are applied in intrusion detection systems (Lata & Indu 2013; Kumar & Veerendranath 2014; Prabha & Sukumaran 2014; Guevara et al. 2016; Kala & Christy 2016). Hameed & Rashid (2014) have used Horspool algorithm as a matching process to distinguished between normal and attack records, the main limitation of choosing Horspool algorithm is they matching is done based on eight keys (that mean they need to build eight bad character tables) which mean the processing time is high, therefore

another matching algorithm is more faster for both processing and matching. All are also applied for biological sequences (Huang et al. 2008; Korhonen et al. 2009; Pandiselvam et al. 2014; Allmer 2016; Manikandan & Ramyachitra 2018), where matching algorithms is applied to DNA sequences in order to diagnose the diseases by searching for the existence of diseased DNA sequence (Tun & Swe 2014; Mane & Pangu 2016; Islam & Talukder 2017) or search for DNA mutations (AlKindhi & Sardjono 2015; Masood & Manjula 2018). These algorithms are applied because they are used to find a specific sequence in the DNA (Rajesh et al. 2010).

The research in DNA sequences shows that the looking for the entire DNA sequences is impossible; therefore looking for special area in DNA is considered (Soram & Khomdram 2010). This done by selecting the number of genes subset that related to specific disease which makes disease diagnosing faster (Huang et al. 2012; Mohammadi et al. 2016; Pavithra & Lakshmanan 2017; Mengdi et al. 2018). On another hand, in IDS the challenge in this field is to select the most relevant feature that lead to increase the performance of intrusion detection system (Al-Jarrah et al. 2014; Eesa et al. 2015; Ambusaidi et al. 2016; Varma et al. 2016). In the past researches, no features selection methods were applied for intrusion detection system that based on DNA sequences idea.

1.4 RESEARCH QUESTIONS

1. What kind of DNA encoding best for anomaly IDS?

2. How far the extraction of STR and matching algorithms can improve the performance of anomaly IDS?

3. How to select the best features in anomaly IDS based on DNA method?

1.5 RESEARCH OBJECTIVES

The aim of this research is to propose DNA sequences for anomaly intrusion detection system, with the following objectives:

1. To propose an enhanced DNA encoding methods for anomaly intrusion detection system (DNA-IDS).

2. To improve DNA-IDS by applying efficient STR extraction, pattern matching and feature selection algorithms.

1.6 RESEARCH SIGNIFICANCE

The high usage of internet by the public and companies led to the importance of computer network security. The growth of the internet technology led to increase the intrusions. Single intrusion of a computer network can lose a large amount of data and may lead to money loss. Therefore, the importance of the present research is to build a system that can monitor the computer and network to reduce and secure these systems from external and internal intruders and to avoid the problems that will occur.

1.7 RESEARCH SCOPE

This research focused; firstly on intrusion detection system concepts, types, and techniques; secondly, on DNA concepts, mutation, and DNA encoding methods; thirdly on the pattern matching algorithms method that are suitable for both anomaly intrusion detection system and biological sequences, and finally on feature selection method that is applied to anomaly intrusion detection system. The selection of DNA encoding method is done based on the concept of minimizing the DNA code that can handle and represent all network traffic attributes values and the availability to represent any new values (inspired by Al_Ibaisi et al. (2008) that used random DNA sequences to represent each network value use less space, and have available places for any new values), where Al_Ibaisi et al. (2008) DNA encoding idea inspired from Adleman (1994) that used random DNA sequences to represented all graph vertex, less quantities of nucleotides, and processed larger graph with the DNA sequences quantities. The performance of the proposed system in this research is evaluated by using both Knowledge Discovery and Data mining (KDDCup 99) dataset and Network Security Laboratory-Knowledge Discovery and Data Mining (NSLKDD) dataset. Finally, the evaluation is based on detection rate, false alarm rate, accuracy,

and the time needed for converting network traffic to DNA sequences and the time for matching.

1.8 RESEARCH METHODOLOGY

This research follows the standard experimental-based research methodology for each objective: identify the problem, propose the algorithms, conduct the experiments and analysis the results for each objective. In order to achieve these objectives, this work is divided into 5 phases: (1) Identify the problem by studying the existing intrusion detection system researches and methods that lead us to identify the weaknesses of those methods; (2) Data collection that review the two datasets that have been used for intrusion detection system evaluation; (3) Propose a DNA encoding for anomaly intrusion detection system (DNA-IDS) and this is done by proposing an enhance three encoding methods based on different DNA characters numbers; (4) Improve DNA-IDS by extracting more STR, applying different matching algorithms, and proposing new features selection method, where the extracting of more STR may lead to enhance FAR result, then applying four matching algorithms to find which one is the most suitable for the proposed method. These algorithms are Brute force algorithm, Boyer Moore algorithm, Horspool algorithm, and Knuth-Morris-Pratt algorithm, and the combination of Boyer Moore Algorithm and Knuth–Morris–Pratt Algorithm. Then, proposing a feature selection method that is depending on DNA location (STR position); (5) calculation and evaluation of the results based on detection rate (DR), false alarm rate (FAR), accuracy, and time (that include both encoding time and matching time). The research methodology details are discussed further in Chapter III.

1.9 THESIS ORGANIZATION

This thesis is divided into seven chapters; the details concern each chapter are summarized as follows:

Chapter II provides the literature review of intrusion detection system, methods that are used for intrusion detection system; DNA components and its analysis methods; describe the methods and applications that are used in this research;

and literature review of feature selection for intrusion detection system and for medical diagnosis.

Chapter III provides the research methodology that is applied to accomplish the research objectives. Such methodology contains multiple phases including the datasets used, the proposed DNA encoding method for anomaly intrusion detection system, improved the proposed method based on STR extraction and matching algorithms, identifies best features, and the performance evaluation.

Chapter IV discussed the components of the proposed method (DNA-IDS), its techniques, and components. The architecture of DNA-IDS is explained, and the results are calculated based on parameters such as DR, FAR, and accuracy. This chapter ends with the proposal of a new DNA encoding method and technique for anomaly intrusion detection system.

Chapter V improved the proposed method DNA-IDS by extracting more STR, applying different matching algorithms, and proposing a new feature selection method. In the first improvement, more STR are extracted to find which DNA encoding will achieve the best results and to enhance the results. In the second improvement, four matching algorithms and combination of two algorithms are applied to find which algorithm is the most suitable for the proposed method that achieved the best result. The results are calculated based on matching time. In the third improvement, proposed a new feature selection method for DNA-IDS, this method is used to select the best features instead of using all 41 features of network traffic records; this process is done by using DNA location. The results are calculated based on parameters such as DR, FAR, accuracy, and execution time. The chapter is ended with the selection of the best DNA encoding method and the best matching algorithm for the DNA-IDS. Also propose of a new feature selection method for DNA-IDS to decrease the execution time.

Chapter VI aims to provide the conclusions of the current study and discuss several suggestions for future works.

CHAPTER II

LITERATURE REVIEW

2.1 INTRODUCTION

The objective of this chapter is to review the state of art for the related fields in this research. It starts with a background of intrusion detection systems in computer system. The literature reviews of existing intrusion detection system approaches and methods; their strengths and weaknesses. Later, this chapter explains the background of DNA in human body and then a details of the concepts of DNA computing and their uses in computer system are exhibited; discussed the biological sequence analysis and the matching algorithms; finally, it showed the feature selection types and methods.

This chapter is organized as following: Section 2.2 describes a background of Intrusion Detection System. Section 2.3 describes several methods used for Intrusion Detection System. Section 2.4 describes DNA components and its analysis, Section 2.5 describes the matching algorithms that used in biological sequences. Finally, Section 2.6 describes the feature selection methods that are used for Intrusion Detection System and for medical diagnosis.

2.2 INTRUSION DETECTION SYSTEM

Security has become a critical issue for modern computer systems due the rapid growth of computer networks during the past two decades (Singh 2004). System security depends on three concepts; these concepts are confidentiality, integrity, and availability. The confidentiality means that the attack causes a confidentiality violation if it allows attackers to access data without authorization (either implicit or

explicit) from the owner of the information. Integrity means that an attack causes an integrity violation if it allows the unauthorized attacker to change the system state or any data residing on or passing through a system. Finally, availability means that an attack causes an availability violation if it keeps an authorized user (human or machine) from accessing a particular system resource when, where, and the form that they need it (Bace & Mell 2001). Attacks can be classified either "Passive" when a network intruder intercepts data traveling through the network, and "Active" in which an intruder initiates commands to disrupt the network's normal operation (Pawar & Anuradha 2015). There are several computer security types such as an antivirus, firewall, and intrusion detection system (IDS).

The concept of intrusion detection expanded by Denning (1987) is the development a model for a real-time intrusion detection system. He presented a method that used a real-time collection of audit records from attempted break-ins, system penetrations, and abuses through applying monitoring system. The used abnormal information categorized into bundles, called tuples, and models are applied to the data. Denning's analysis detected a wide range of intrusions. Some of the detected intrusions are identified without the knowledge of system vulnerabilities.

Intrusion is any set of actions that attempt to violate the integrity, confidentiality, or availability of a computer resource (Sazzadul & Bikas 2012). IDS is a software application or device that used to monitor the network or system for malicious activities and send report to management station (Scarfone & Mell 2007). The aim of IDS is to detect the attacks that happen in computer and network. IDS can detect both internal and external attacks. Figure 2.1 shows the possible locations of intrusion detection system.

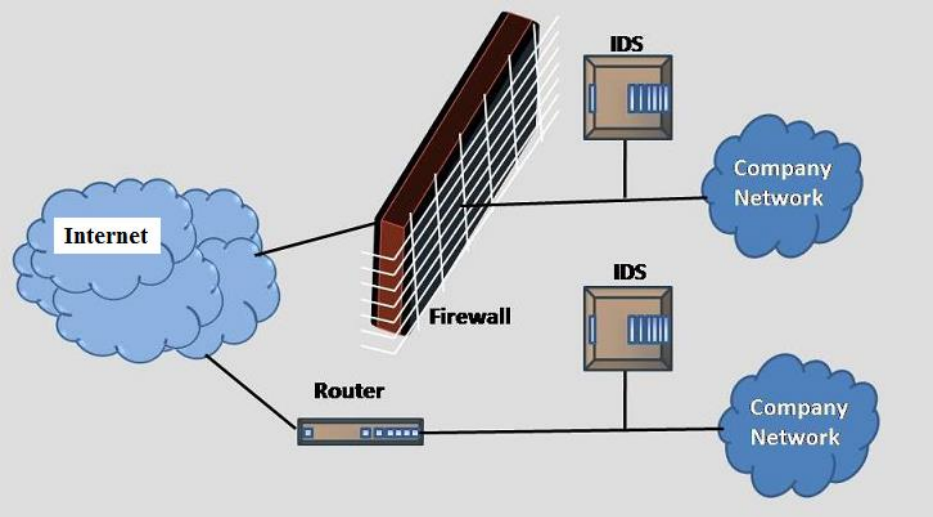

Figure 2.1 Possible locations of IDS (Gandhi & Srivatsa 2010)

The development of IDS is motivated by the following factors:

- Most existing systems have security flaws that render them susceptible to intrusions. Finding and fixing all these deficiencies are not feasible (Denning 1987) and in particular, complex systems are prone to errors that could be exploited by malicious users.

- Prevention techniques are not sufficient. It is almost impossible to have a secure system (Denning 1987). IDS appears when an intrusion has occurred and cannot be prevented by existing security systems.

- In the early 2000s, a new threat like SQL injections become popular and this attack would pass right by the firewall. Hence, this is the real beginning of putting the IDS into use (Pirc 2017).

- In the period between 2006 – 2010, security companies offered IDS solutions using more Gbps in order to provide the ability to monitor more segmented networks (Pirc 2017).

- The years between 2011 and 2015, huge turning points for IDS vendors were introduced, as they began creating next generation, which included features such as application and user control (Pirc 2017).

- Since authorized users generally considered as insider threats, even the most secure systems are susceptible to insiders. Furthermore, many organizations expressed that threats from inside can be much more harmful than outsider attacks.

- New intrusions continually emerge. Therefore, security solutions need to be improved or introduced to defend systems against novel attacks. This is what makes intrusion detection such an active research area.

- Until todays, intrusion detection is changing and will likely continue to change as threat actors change the tactics and techniques they use to break into networks (Pirc 2017).

Intrusion detection system consists of three components; these components are data collection, detection, and response. Data collection is used to collect and pre-process the data, such as transform the data to specific format, storage of data, and sending the data to detect module (Lundin & Jonsson 2002). The detection module will analyze and process the data obtained from the data collector model in order to detect intrusion attempts, and then forward the events flagged as malicious to the response module. The response behaves according to the response policy defined, and these responses are divided into active and passive responses (Axelsson 2000). Figure 2.2 shows the general intrusion detection system components.

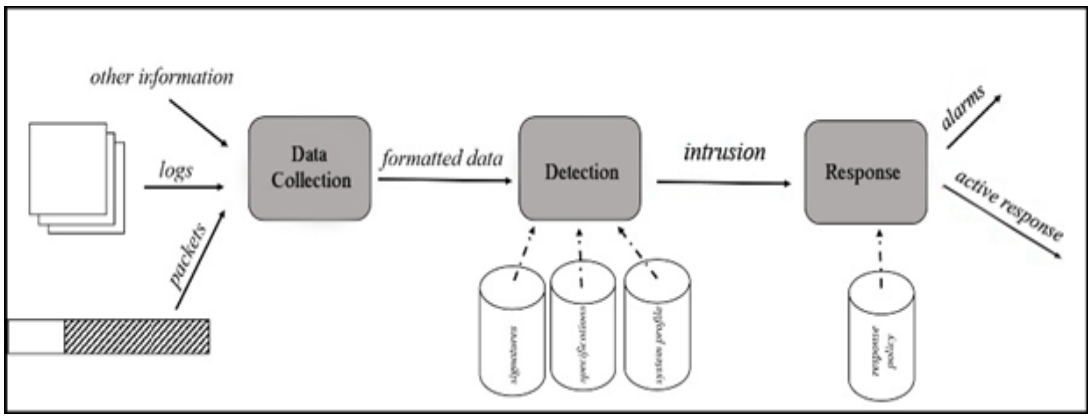

Figure 2.2 IDS components (Sen 2015)

An IDS has several advantages. It can detect both internal and external attackers, provides easy protection system for the entire network, provides centralized management, and provides an additional layer of protection. On the other hand, IDS has some disadvantages, IDS generates large number of alarms that increase analysis workload, most of these alarms are false alarms, required high performance of an IDS, and required large training data to characterize normal behavior patterns (Meng & Li 2012). An IDS can be categorized in different ways. The major categorizations are shown in Figure 2.3.

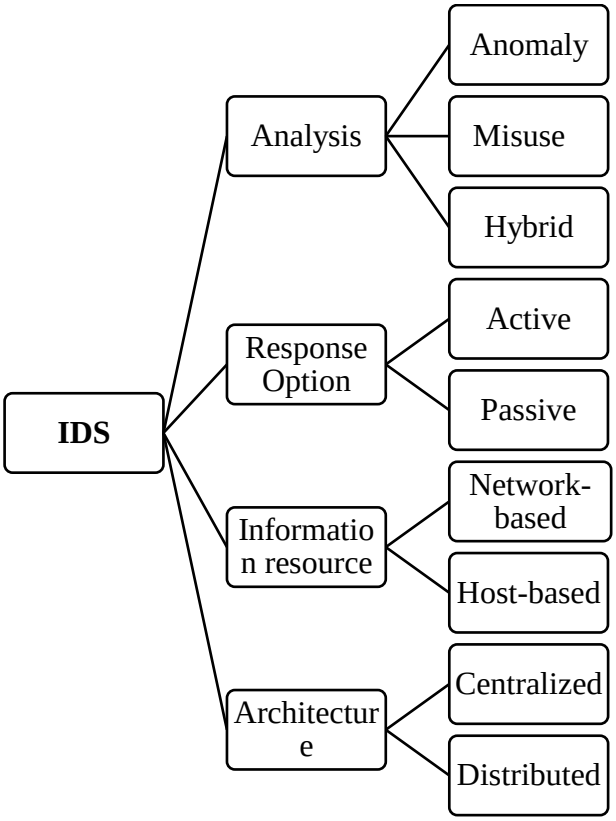

Figure 2.3 Classification of IDS (Debar et al. 1999)

As mentioned above, intrusion detection system can be classified as passive or active system. In a passive system, IDS sensor detects something that looks like a gap in the system and signals an alert but does not take any prevention method to stop the attack. Active IDS responds to the suspicious activity in the same manner as the passive IDS with the added ability to take action on the attack (Kemmerer & Vigna 2002). IDS can be classified as Network Intrusion Detection Systems (NIDS) or Host-based Intrusion Detection Systems (HIDS), based on the information sources that they are used. Network Intrusion Detection Systems analyze network packets captured

from a network segment, and these systems use software programs called sensors to collect network packets. While Host-based Intrusion Detection Systems examine audit trails or system calls generated by individual hosts. This system has difficulty to detect attacks, since it monitors only information gathered from the computer system (Endorf et al. 2003). Intrusion detection system can also be categorized according to the detection approaches they used. There are two detection methods: misuse detection and anomaly detection. The misuse detection identifies intrusions based on features of known attacks while anomaly detection analyzes the properties of normal behavior (Zhang & Zulkernine 2006). Finally, intrusion detection system can also be categorized according to the architecture as Centralized or Distributed intrusion detection system. A centralized IDS data is sent to central location independently from the target before it is analyzed; this location is an analysis engine (Kozushko 2003). A Distributed IDS consists of multiple IDS over a network which communicates with each other, and can get a view of what happen in the whole network (Einwechter 2001).

2.3 METHODS FOR INTRUSION DETECTION SYSTEM

Over the last few years, intrusion detection systems have been widely discussed. Earlier, IDS developed based on sampling sets on periodic basis as the first method that used to monitor the security of computer systems (Anderson 1980). Later, Denning (1987) developed a model for a real time IDS. Since that, various approaches have been proposed for IDS solutions, the most proposed approaches are: Artificial Neural Network, Bio-inspired Computing, Evolutionary Technique, Machine Learning, Pattern Recognition, and these are shown in Figure 2.4 below.

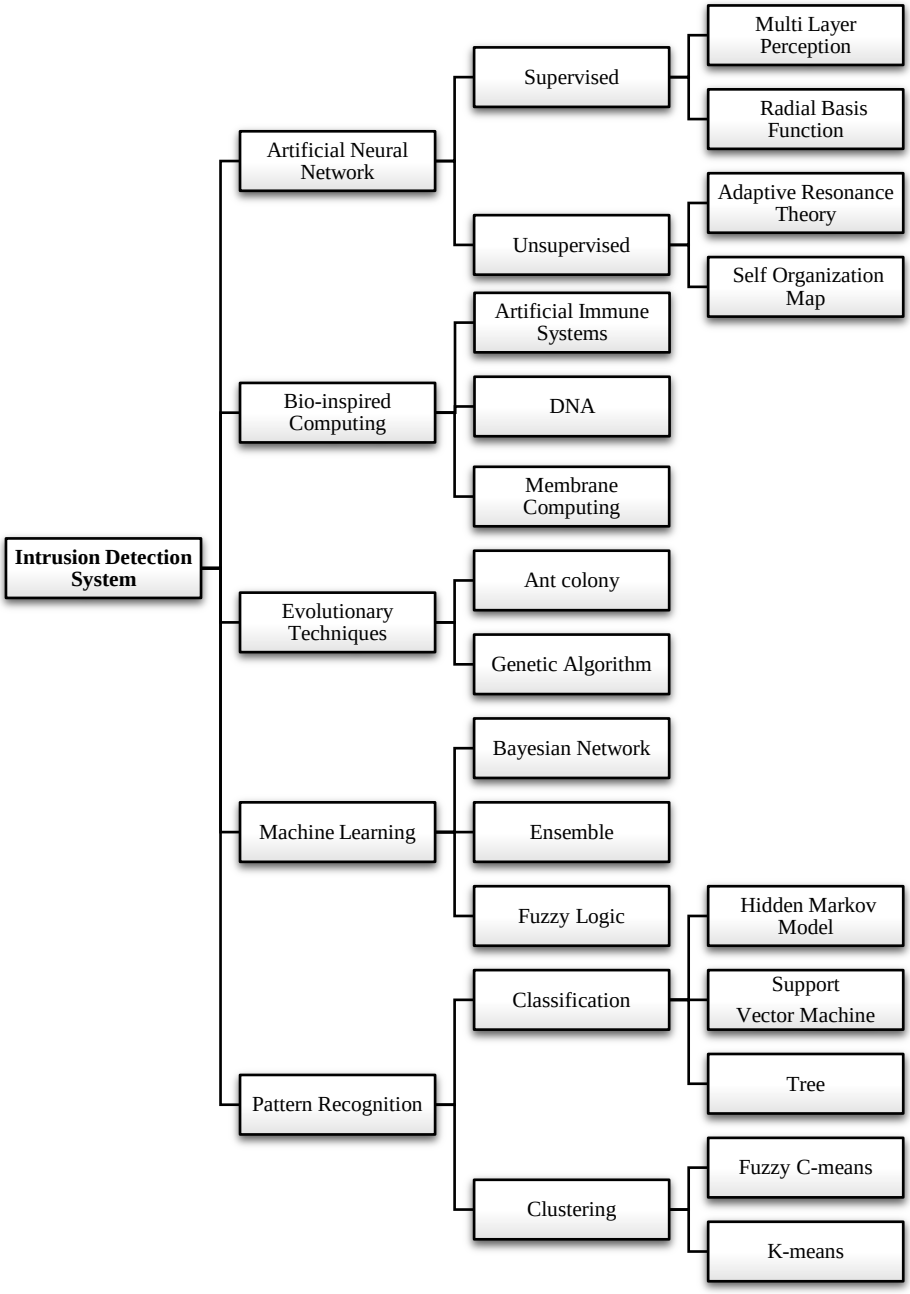

Figure 2.4 Taxonomy of intrusion detection system

2.3.1 Artificial Neural Networks for IDS

Artificial neural network is an information processing model that is inspired by the biological nervous systems, such as brain. It tries to represent the physical brain and thinking process through electronic circuit or software. Artificial neural network is the network of individual neurons. Each neuron is a neural network acts as an independent

processing element. Each processing element (neuron) is fundamentally a summing element followed by an activation function. The output of each neuron is fed as the input to all of the neurons in the next layer (Theodorios & Koutroumbas 1999).

Artificial Neural Networks have been used in many research fields such as: medicine, biology, engineering, and other. Artificial neural networks are used to find a solution for some problems, one of these problems is a pattern recognition problem. One of the first attempts for intrusion detection based on neural network was proposed by Ryan et al. (1998) that built offline intrusion detection system for ten users by using three layers architecture.

Many studies of artificial neural network have been proposed for intrusion detection system, such as (Haddadi et al. 2010; Norouzian & Merati 2011; Shi et al. 2011; Barman & Khataniar 2012; Wang & Yu 2013; Hu et al. 2015; Ingre & Yadav 2015; Xing-zhu 2016; Azad & Jha 2017; Chun et al. 2017; Kim et al. 2017). Some of these systems are based on one method of artificial neural network, such as a system proposed by Al-Janabi and Saeed (2011), where they used back propagation artificial neural network to learn system's behavior, and a system developed by Poojitha et al. (2010) based on Feed Forward Neural Network. Some approach for intrusion detection is proposed based on combination of two systems, such as a system proposed by Wang et al. (2010) that combines artificial neural networks and fuzzy clustering, and the intrusion detection system that built based on the combination of k-means clustering and artificial neural network (Tammi et al. 2015). There are many disadvantages of applying artificial neural network for intrusion detection system such as it is not suitable for real time because the training time is slow, and the whole system must be retrained when added new attack. Table 2.1 shows the reviews of some intrusion detection systems using artificial neural network technique.

Table 2.1 Reviews for some IDS based on artificial neural network techniques

Author(s)	Year	Method(s)	Dataset	Performance
Haddadi et al.	2010	Feed-forward neural network	KDDCup 99	DR results for normal records, DoS attack, probe attack, R2L attack, and U2R attack are equal to 79.8%, 97.5%, 99.1%, 98.9%, and 34.5% respectively.
Wang & Yu	2013	Radial basis Functions and Elman	1998 DARPA	For anomaly detection system: the DR and FAR results are equal to 93% and 2.3% respectively. For misuse detection system: the DR and FAR results are equal to 95.3% and 1.4% respectively.
Xing-zhu	2016	Radial basis Functions	KDDCup 99	Used 1240 instances for testing, best DR results for DoS attack, probe attack, R2L attack, and U2R attack are equal to 95.80%, 93.95, 95.20%, and 96.02% respectively.
Azad & Jha	2017	Fuzzy min max neural network	KDDCup 99	Results are good as compared to the other well-known methods.
Chun et al.	2017	Neural Network	NSL-KDD	The obtained DR results for DoS attack, Probe attack, and R2L attack are equal to 82.39%, 81.16%, and 20.8% respectively. The achieved DR and accuracy results are equal to 83.42% and 80.34% respectively.
Kim et al.	2017	Deep Neural Network	KDDCup 99	DR and accuracy results are equal to 99%, and FAR result is equal to 0.08%.

2.3.2 Bio-inspired Computing for IDS

There is a similarity between computers and organisms. Bioinformatics is the modeling of biological processes using the information technology; bioinformatics have helped to know these similarities more accurately and can use for computer security (Goel & Bush 2004). Intrusion detection system problem can be solved based on Bio-inspired such as Immune Systems, DNA, and Membrane Computing.

The natural human immune system defends the human body against harmful and previously unseen foreign cells using lymphocyte cells. The foreign cells are called antigens, such as bacteria and viruses (De Castro & Timmis 2002). The artificial immune system research began in the 1986 by Farmer et al., (1986) when

they suggested that the immune system could be used with computer science. The artificial immune system is designed for the computational system and inspired by the human immune system. This system incorporates many attributes of the human immune system including diversity, error tolerance, dynamic learning, and adaption and self-monitoring (Hofmeyr & Forrest 2000). It applied to solve various problems in the field of information security (Harmer et al. 2002). A classical theory (Forrest et al. 1994) that proposed the immune system for computer security is considered as a measure for intrusion detection system.

Numerous studies of artificial immune system have been proposed for intrusion detection system, such as (Zhang et al. 2011; Hosseinpour et al. 2014; Abas et al. 2015). A method for network intrusion detection based on an artificial immune system and Holland's Classifier was proposed by Randrianasolo & Pyeatt (2012) based on a network with 50 nodes and 50 intruders. A new system is built by Kumar and Reddy (2014) to gather information from various nodes. This information is used with an evolutionary artificial immune system to detect the intrusion and they proved that the proposed system is well suited for intrusion detection and prevention in wireless network.

Another Bio-inspired is DNA; the DNA is the genetic material found in most organisms (including human). Every computer system can be given a DNA characterization (Yu et al. 2001) that include sequences or genes responsible for characteristics such as network traffic, system calls, and user behavior. Some research of DNA have been proposed for intrusion detection system, such as (Mahdy & Saeb 2007; Al-Ibaisi et al. 2008; Hameed & Rashid 2014). There are many factors effect on DNA system; these factors are DNA encoding, DNA's Keys and their positions, and matching process. DNA encoding is the process of converting plaintext to DNA sequence, DNA encoding is applied in different fields of computer system such as cryptography, steganography, digital signature, and other. An anomaly intrusion detection system is introduced by Mahdy & Saeb (2007), using the concept of a DNA sequence or gene, which is responsible for the normal network traffic patterns. Then, the system detects suspicious activities by searching the "normal behavior DNA sequence" through string matching. This system is performed by using two steps: the

first step is the network traffic DNA sequence generation or the offline training phase, and the second step is the network monitoring or the online monitoring phase, where a monitoring phase is adopted, and the hardware is implemented with DNA pattern matching is performed.

A normal signature sequence and alignment threshold value generated from processing the system training data and encoded network connection into corresponding DNA nucleotides sequence is adopted (Al-Ibaisi et al. 2008). Then, the signature sequence was aligned to find the similarity degree value and to decide whether the connection was an attack or normal. New generations are produced to choose the signature with best alignment value with normal network connection sequences. This system is carried out by applying two steps: the first step included the DNA sequence encoding, where the main idea of their encoding is less space and have available nucleotides for any new values. Also, it divided the network traffic values either static parameters or dynamic parameters, where the static parameters included flag, protocol, and services, while the dynamic parameters included integer, real, and Boolean. In the second step, the genetic algorithm is used to enhance the selection for target solution. To enhance their system results, one of their suggestions is to use a new DNA encoding idea.

A misuse intrusion detection system based on DNA sequence has been introduced by Hameed & Rashid (2014). Three steps are followed to perform this system. In the first step, enhanced the DNA encoding method that proposed by Al-Ibaisi et al. (2008) by using the same value to represent both integer and Boolean attributes and then used two characters to represent these values instead of three characters that was used previously. This encoding procedure is used to convert the network traffic to DNA sequences. In the second step; the attack signature keys and their positions are extracted by using Teiresias algorithm. In the last step, the network traffic classified as either attack or normal based on attack signature keys and their positions by using Horspool algorithm. They used 4000 random records from 10% KDD dataset for both training and testing phases.

Another Bio-inspired is membrane computing. Membrane computing is a bio-inspired branch of natural computing, abstracting computing models from the structure and functioning of living cells and from the organization of cells in tissues or other higher order structures (Păun 2010). Some researches of membrane computing have been proposed for intrusion detection system, such as (Idowu et al. 2013, Idowu et al. 2014, and Idowu et al. 2015). Idowu et al. (2013) proposed a new algorithm to solve the NP complete optimization problem based on P-system paradigm, and also outlined the preliminary achieved results of using membrane computing to enhance Bee algorithm for anomaly IDS. A hybrid method based on membrane computing and Bee algorithm for anomaly IDS is also proposed by Idowu et al. (2014) with a view to reduce useless features, the achieved results showed a high detection rate and decreased the false positives/negatives. After that, Idowu et al. (2015) suggested the membrane computing method that used as a tool to solve some network security systems problems such as packet dropping in Intrusion Detection System (IDS). The initial achievement results exhibited the speedup of the detection.

Table 2.2 shows the reviews of some intrusion detection systems using Bio-inspired methods.

Table 2.2 Reviews for some IDS based on bio-inspired methods

Author(s)	Year	Method(s)	Dataset	Performance
Mahdy & Saeb	2007	DNA sequence and pattern matching	Collected data	Matching is implemented on FPGA, and a fast searching process is achieved.
Al-Ibaisi et al.	2008	DNA sequence encoding and Genetic algorithm	KDDCup 99	The obtained DR results for DoS attack, Probe attack, R2L attack, U2R attack and normal records are equal to 51.83%, 57.28%, 24.20%, 43.10%, and 79.79% respectively.
Hosseinpour et al.	2010	Artificial Immune System and GA	-	Decrease the detection time for each connection.
Zhang et al.	2011	Artificial Immune System	NSL-KDD	The achieved DR results for normal records, DoS attack, Probe attack, and R2L attack are equal to 96%, 96.04%, 89.8%, and 99.7% respectively.

to be continued…

…continuation

Randrianasolo & Pyeatt	2012	Artificial Immune System and Holland's classifier	-	The achieved results, for DR is equal to 90.57%, for false positive rate is equal to 17.21%, and for false negative percentage is equal to 9.42%.
Idowu et al.	2013	Membrane Computing	KDDCup 99	The achieved DR results are equal to 80.62%, 93.07%, and 93.63%. While the achieved FAR results are equal to 0.009%, 0.001%, and 0.001%.
Hameed & Rashid	2014	DNA sequence encoding	KDDCup 99	Used 4000 instances for testing, The DR, FAR, and accuracy obtained values by using keys only are equal to 99%, 27.2%, and 94.4% respectively. While, when keys and their positions are used, the values are equal to 97.57%, 1%, and 97.82% respectively.
Hosseinpour et al.	2014	Artificial Immune System	KDDCup 99	Used 22545 records, the obtained results for FPR is equal to 0.8%, for TNR is equal to 99.1%, and accuracy value is equal to 77.1%.
Idowu et al.	2014	Membrane Computing	KDDCup 99	The mean for achieved DR and FAR results are equal to 89.11% and 0.004% respectively.
Kumar & Reddy	2014	Artificial Immune System	-	Proved that the system is suited for intrusion detection and prevention in wireless network
Abas et al.	2015	Artificial Immune System	GureKddcup	They obtained high TP and accuracy values, and less time.
Idowu et al.	2015	Membrane Computing	KDDCup 99	The achieved results show that the proposed system is speedup the detection.
Chiba et al.	2018	Back Propagation	KDDCup 99	Achieved high DR, accuracy, and F-score results, and low FPR.
Farahnakian & Heikkonen	2018	Deep Auto Encoder	KDDCup 99	The achieved DR, FAR, and accuracy results are equal to 95.65%, 0.35%, and 96.53% respectively.

2.3.3 Evolutionary Techniques for IDS

Evolutionary computation technique is a subfield of artificial intelligence inspired from natural evolution. It has been successfully applied to many research areas such as computer networks, software testing, medicine, and art. The most studied area in the security domain was the intrusion detection, and various intrusion detection techniques already exist in the literatures. The following characteristics of

evolutionary computation attract researchers to investigate these techniques on intrusion detection. These are generating readable outputs by security experts, ease of representation, and producing lightweight solutions. Furthermore, evolutionary computation does not require assumptions about the solution space (Fogel 2000).

One of the most popular methods of evolutionary techniques is genetic algorithm. The genetic algorithm was introduced initially for the computational biology field, which uses computer for selection and evolution processes. The algorithm started with generating random candidate programs population, and then used a fitness measure to evaluate each individual performance. Crosbie and Spafford tried to integrate genetic algorithm with intrusion detection system (Crosbie & Spafford 1995). There are many intrusion detection system researches that based on genetic algorithms. Padmadas et al. (2013) proposed intrusion detection based on genetic algorithm by using four layers based on attacks groups. Thamilarasu (2015) provided an intrusion detection system based on multi-objective genetic algorithm to provide attack detection in these networks. Genetic algorithm is used for intrusion detection system to find a chromosome evaluations function to get intrusion detection system solution (Bhattacharjee et al. 2017). However, the main limitation is that incorrect threshold value might lead to high FAR results and time-consuming (Ghorbani et al. 2010). Table 2.3 shows the reviews of some intrusion detection systems using evolutionary techniques.

Table 2.3 Reviews for some IDS based on evolutionary techniques

Author(s)	Year	Method(s)	Dataset	Performance
Li et al.	2011	Ant Colony and Fuzzy C-means	KDDCup 99	The obtained DR values for Dos attack, probe attack, R2L attack, U2R attack, and normal records are equal to 95.34%, 90.03%, 13.82%, 34.18%, and 99.41% respectively.
Cai & Yuan	2013	Ant Colony	KDDCup 99	Reduce computation time, precision and recall values are equal to 96.94% and 98.41% respectively.
Padmadas et al.	2013	Genetic Algorithm	Collected data	Achieved an accuracy result for R2L attack equal to 90%.

to be continued…

…continuation

Abdurrazaq et al.	2014	Ant Colony	KDDCup 99	Used 1000 records for testing, these records are either normal records or probe attack records. Achieved high detection performance in real-time.
Desale & Ade	2015	Genetic Algorithm	NSL-KDD	Achieved high accuracy and reduced time.
Thamilarasu	2015	Genetic Algorithm	Collected data	Decreasing the computational complexity.
Qiang et al.	2016	Ant Colony	KDDCup 99	The DR values for DoS attack, probe attack, R2L attack, and U2R attack are equal to 93.1%, 92.2%, 90.7%, and 95.5% respectively.
Varma et al.	2016	Ant Colony	UCI Cleveland	Accuracy value is increase by 0.11% after the reduction of features, and time is faster by 37.19% compared to all features.
Bhattacharjee et al.	2017	Genetic Algorithm	NSL-KDD	Fuzzy Vectorised GA gave better detection results than the Vectorised GA and Weighted Vectorised GA.
Tabatabaefar et al.	2017	Particle Swarm Optimization	KDDCup 99	DR result is equal to 99.1%, FAR result is equal to 1.9%, and accuracy result is equal to 99.58%.

2.3.4 Machine Learning for IDS

Several anomaly intrusion detection systems are proposed based on different machine learning techniques. Some systems are based on single learning techniques such as Bayesian Network, and Fuzzy Logic. Other systems are based on the combining of different learning techniques, such ensemble techniques (Tsai et al. 2009). The main drawback is the high consumption of resources.

The Bayesian network is used as a graphical representation over a set of variables. The Bayesian networks have been used in different computer science fields (Darwiche 2010) because it has the ability to obtain result from probabilistic information. There are many efficient algorithms that can be used to obtain results from the information. One can build an efficient intrusion detection system based on these algorithms (Altwaijry & Algarny 2012). Fuzzy logic is a form of many-valued logic that deals with approximate, not fixed and exact, and has been used in computer and network security area since 1993 (Hosmer 1993). Fuzzy method can be applied to intrusion detection system because it can consider the features as fuzzy variables. The

main disadvantages of applying fuzzy logic are the high resource consumption (Garcia-Teodora et al. 2009). Another machine learning technique is ensemble learning that uses and combines various classifiers and each classifier can search in different areas and then combine these search results, this lead to a better learning than a single classifier (Zhao 2013). The most commonly used ensemble algorithms are bagging and boosting.

There are many intrusion detection system researches based on machine learning methods for example Koc et al. (2012) applied Hidden Naïve Bayes method to intrusion detection system that improved DoS attack accuracy. Modil et al. (2012) designed and integrated the Bayesian classifier and Snort to build a network intrusion detection system in the cloud. The aims of the framework were to detect network intrusions in cloud environment with low false positives and with affordable computational cost. Falke et al. (2014) designed an intrusion detection system based on fuzzy logic in order to monitor the intrusion activities. An intrusion detection system implemented by Gaikwad and Thool (2015) based on bagging ensemble method reduced the features number to 15 features from 41 features. Sreenath and Udhayan (2015) used the Bagging Ensemble Selection to establish a novel method for intrusion detection; the outcomes showed that the developed method is highly effective to overcome the drawbacks found in previous works. Belavagi and Muniyal (2016) proposed an intrusion detection system by using different machine learning algorithms and they compared between their performance results. Jabbar et al. (2017) proposed an intrusion detection system based on ensemble classifier that used both Random Forest algorithm and Average One-Dependence Estimator algorithm. Mkuzangwe and Nelwamondo (2017) presented an intrusion detection system based on fuzzy logic method to detect Neptune attack. A new visualization tool for intrusion detection system is introduced by Theron et al. (2017) based on the combination between Principal Component Analysis (PCA) and Group-wise PCA, these lead to a tool that is highly flexible and also allow the user to move between huge amounts of data to find unexpected behaviors. The drawbacks of applying artificial immune system are playing for intrusion detection system such as complexity and high false positive rates results. Table 2.4 shows the reviews of some intrusion detection systems using machine learning methods.

Table 2.4 Reviews for some IDS based on machine learning methods

Author(s)	Year	Method(s)	Dataset	Performance
Altwaijry & Algarny	2012	Naive Bayesian	KDDCup 99	The obtained DR values for Dos attack, probe attack, R2L attack, U2R attack, and all attacks are equal to 99.36%, 57.17%, 0%, 0%, and 89.70% Respectively.
Devarakonda et al.	2012	Bayesian Network and Hidden Markov Model	KDDCup 99	The model classification result is high.
Koc et al.	2012	Hidden Naïve Bayes	KDDCup 99	The obtained accuracy result is equal to 93.27%, and the error rate result is equal to 6.28%.
Modi et al.	2012	Bayesian Classifier and Snort	KDDCup 99	Achieved high accuracy result that is equal to 97.07%, and base rate result equals to 77.06%.
Li & Lin	2013	Rough Classifiers	1999 DARPA	FAR result is equal to 10.40%, and DR for DoS attack, probe attack, R2L attack, and U2R attack are equal to 86.25%, 89.56%, 73.49%, and 76.67% respectively.
Falke et al.	2014	Fuzzy Logic	KDDCup 99	High system accuracy and an efficient intrusion detection.
Balan et al.	2015	Fuzzy Logic	-	Provided a secure communication between nodes and detected attacks efficiently.
Chapke & Deshmukh	2015	Fuzzy Logic and C4.5	KDDCup 99	The obtained DR and FAR results are equal to 99.47% and 2.75% respectively.
Gaikwad & Thool	2015	Bagging method	NSL-KDD	The DR, FAR, and accuracy result for specific features are equal to 78.4%, 17.2%, and 78.37% respectively.
Sreenath & Udhayan	2015	Bagging method	NSL-KDD	The obtained accuracy value is equal to 97.85%.
Farnaaz & Jabbar	2016	Random Forest	NSL-KDD	The best DR result is equal to 99.84%, and accuracy value is equal to 99.67%.
Li et al.	2016	Extreme Learning Machines	KDDCup 99 and NSL-KDD	Used three types of attacks: DoS, probe, and R2L, and they used each attack separately. ELM gave better results than SVM
Rodda & Erothi	2016	Machine learning	NSL-KDD	The obtained DR results for DoS attack, probe attack, R2L attack, and U2R attack are equal to 95.1%, 98.13%, 93.35%, and 19.04% respectively.

to be continued…

… continuation

Jabbar et al.	2017	Ensemble classifier	Kyoto	The DR result is equal to 92.38%; FAR result is equal to 0.14%, and accuracy value is equal to 90.51%.
Mkuzangwe & Nelw Amondo	2017	Fuzzy Logic	NSL-KDD	The obtained accuracy result is equal to 93.24%.
Theron et al.	2017	PCA	UGR'16	High detection accuracy and diagnosis of abnormal network behavior.
Belouch et al.	2018	Machine learning	UNSW-NB15	The achieved DR and accuracy results are equal to 97.49% and 93.53% respectively.
Idhammad et al.	2018	Naive Bayes	CIDD-001	The achieved FAR and accuracy results are equal to 0.21% and 97.05% respectively
Verma & Ranga	2018	K-nearest neighbor	CIDD-001	The achieved accuracy results with different number of neighbors are equal to 95%, 96%, 94%, 95%, and 93%.

2.3.5 Pattern Recognition for IDS

The general aim of the pattern recognition algorithms is to provide a reasonable answer for all possible inputs and to perform matching of the inputs, considering their statistical variations. An example of the pattern matching algorithm is the regular expression matching, which looks for patterns of a given sort in textual data and is included in the search capabilities of many text editors and word processors *(*Bishop 2011). Algorithms for pattern recognition are depending on the type of label output, on whether learning is supervised (Classification algorithms) or unsupervised (Clustering algorithms). Classification is applied in various applications that deal with a huge data, such as plaintext classification, medical diagnosis, intrusion detection system, and other. Different classification techniques used in intrusion detection system such as decision tree, support vector machine and others.

Numerous studies of pattern recognition methods have been proposed for intrusion detection system, such as (Ganapathy et al. 2012; Karthick et al. 2012; Eslamnezhad & Varjani 2014; Gupta & Kulariya, 2016). Wang (2011) improved K-means clustering for intrusion detection method that trains on unlabeled data in order to detect new attacks. Yang and Lin (2015) improved Sunday algorithm by added an

extra jump array and applied it for intrusion detection system. Chakir et al. (2016) proposed a new alert classification algorithm based on pattern matching algorithm for intrusion detection system. Chen et al. (2016) proposed a state-based Hidden Markov Model classification method to detect the advanced attacks with a sequence of attack stages. Dagar et al. (2016) applied different pattern matching algorithms for intrusion detection system, these algorithms are Naïve algorithm, Rabin Karp Algorithm, Knuth-Morris Pratt algorithm, and Tree data structure, then calculated the efficiency of these algorithms based on time. Farnaaz and Jabbar (2016) built an intrusion detection system based on random forest classifier, and then compared the results with traditional classifiers results. Yin et al. (2016) proposed a novel intrusion detection system based on support vector machine and context validation, the context validation is used as preliminary analysis to remove noise that may happen by false alarm. Aldwairi et al. (2017) suggested an intrusion detection system based on pattern matching algorithm, and applied Myers algorithm with different frameworks in order to speed up the matching process.

Chakir et al. (2017) proposed a new alert management system based on pattern matching, it used for intrusion detection system and can classify alerts depending on their importance. An intrusion detection system based on multi class support vector machine proposed by Ikram and Cherukuri (2017) that lead to reduce the time for both training and testing. They used chi-square method for feature selection. Matkar et al. (2017) proposed new algorithms based on pattern matching algorithms that have been used widely in intrusion detection system by analyzing the advantages and disadvantages of these algorithms. Rajasekaran and Ayyasamy (2017) built a novel intrusion detection system based on the combination of support vector machine classifier, k-nearest neighbor classifier, and attribute selection algorithm, then used Incremental Particle Swarm Optimization to enhance the classification accuracy. Hybrid intrusion detection system proposed by Tewatia and Mishra (2017) where pattern matching algorithm applied for misuse detection while the clustering algorithm applied for anomaly detection. A new clustering method proposed to solve intrusion detection system problem through calculating the distance between two samples only one time (Wei et al. 2017). Table 2.5 shows the reviews of some intrusion detection systems using pattern recognition methods.

Table 2.5 Reviews for some IDS based on pattern recognition methods

Author(s)	Year	Method(s)	Dataset	Performance
Wang	2011	K-means	KDDCup 99	Proved that the improved k-means system is given higher DR result and lower FP result than k-means algorithm.
Ganapathy et al.	2012	Fuzzy C-Means and Immune genetic algorithm	KDDCup 99	Average of recall and precision results are equal to 94.16% and 94.86% respectively. FCM based on Immune genetic algorithm gave better FAR result than FCM only.
Karthick et al.	2012	Hidden Markov Model	1999 DARPA	Achieved best accuracy result, it is equal to 97.1%, and best FAR result that is equal to 2.71%.
Modil et al.	2012	Bayesian Networks	KDDCup 99	The best performance is achieved when base rate was equal to 74.68%. The best TP and accuracy results for this base rate are equal to 96.57% and 97.07% respectively.
Mohammad & Sulaiman	2012	SVM	Collected data	Used 55000 records for testing, the best accuracy result is equal to 99.60%, and CPU run time is equal to 15.44 seconds.
Eslamnezhad & Varjani	2014	MinMax K-means	NSL-KDD	MinMax K-means algorithm has higher DR and lower FP results than K-means algorithm results; and these results are equal to 81% and 9% respectively.
Kim et al.	2014	C4.5 Decision Tree	NSL-KDD	Detection time for training and testing are equal to 21.37 seconds and 10.13 seconds respectively.
Fossaceca et al.	2015	Extreme Learning Machines	KDDCup 99	Test dataset has 72793 instances, DR results for DoS attack, probe attack, R2L attack, and U2R attack are equal to 99.96%, 97.42%, 94.94%, and 62.87% respectively.
Kao et al.	2015	Pattern Matching	-	They achieved 86% of the performance and reduce the size by 20% compared to the original algorithm.
Yang & Lin	2015	Pattern Matching	-	Improved the matching efficiency, and it's faster than the original algorithm.
Chakir et al.	2016	Pattern Matching	KDDCup 99	Classify alerts correctly and reduce false alerts.

to be continued…

… continuation

Chen et al.	2016	Hidden Markov Model	Collected data	Accuracy value is equal to 86.2%, precision result is equal to 93.2%, and recall result is equal to 84.1%.
Dagar et al.	2016	Pattern Matching	-	KMP algorithm achieved the fastest time when few amounts of pattern are used, while Tree algorithm is achieved the fastest time when large amount of pattern is used.
Yin et al.	2016	SVM and context validation	KDDCup 99	The training time is 522 seconds. Recall, accuracy, and precision results are equal to 98.40%, 94.16%, and 94.79% respectively.
Aldwairi et al.	2017	Pattern Matching	-	Time results showed an improvement for Phoenix++ and MAPCG MapReduce implementations by 1.3% and 1.7% respectively.
Chakir et al.	2017	Pattern Matching	KDDCup 99	Determine the attacks risk and gave the priority based on their importance, and also reduce FAR value.
Ikram & Cherukuri	2017	SVM and chi-square	NSL-KDD	Training and testing time are equal to 10, and 235 seconds, the highest accuracy value is equal to 98.1%, and the best FAR result is equal to 1.9%.
Lee & Yang	2017	Pattern Matching	-	Achieved 58% higher results than head-body matching algorithm.
Matkar et al.	2017	Pattern Matching	-	Reduced comparisons number that lead to increase the matching efficiency.
Rajasekaran & Ayyasamy	2017	SVM, k-nn, and IPSO.	KDDCup 99	Achieved highest accuracy result when it is compared with the existing methods (single classifiers, ensemble approach, and hybrid approach).
Tewatia & Mishra	2017	Pattern Matching and clustering	-	Merge the benefits of anomaly and misuse detection that led to improve the performance, and converted the unknown intrusion to known intrusion, therefore, improved the accuracy.
Wei et al.	2017	Intra-class distance	KDDCup 99	Used four testing sets, each set contains 3000 records, the training time is equal to 92 seconds, the DR and FAR results are equal to 98.75% and 16.88% respectively.
Yang et al.	2017	Fuzzy interpolation	NSL-KDD	DR results for DoS attack, probe attack, R2L attack, and U2R attack are equal to 98.15%, 74.11%, 75.99%, and 45.71% respectively. The accuracy value is equal to 74.41%.

to be continued…

…continuation

Yuan et al.	2017	C5.0 method and Naive Bayes algorithm	KDDCup 99	The obtained DR results for DoS attack, probe attack, R2L attack, and U2R attack are equal to 97.33%, 87.74%, 12.71%, and 51.43% respectively, FAR result is equal to 6.44%, and accuracy result is equal to 93.32%.
Kabir et al.	2018	Least Square Support Vector Machine	KDDCup 99	Achieved high DR results and low FAR results.

2.4 DEOXYRIBONUCLEIC ACID

DNA is the genetic material found in most organisms (include human) as the main chromosomes component which is used to transfer the genetic information (Soram & Khomdram 2010). The main advantage of DNA is the storage of information. The information in DNA is stored as a code made of four chemical bases, these bases called Adenine, Cytosine, Guanine, and Thymine, and referred as A, C, G, and T, to form base pairs that attached to a sugar molecule and a phosphate molecule as shown in Figure 2.5 (Marieb & Hoehn 2006). The order or sequence of these bases makes individual DNA unique and determines the information available for building and maintaining an organism (Soram & Khomdram 2010).

Nucleotides is a base pair with sugar and phosphate, which is form two spiral long strands connected together based on base pairs. There are about 3 million bases, 99% of these pairs are the same with all persons, and only 1% is unique. DNA cells contain genetic information and this information is shared through chromosomes. There is a total of 46 chromosomes, 23 from father and 23 from mother. DNA in human is shared 99.7% with their parents and only 0.3% is the unique code (repetitive coding) that serves for DNA biometrics.

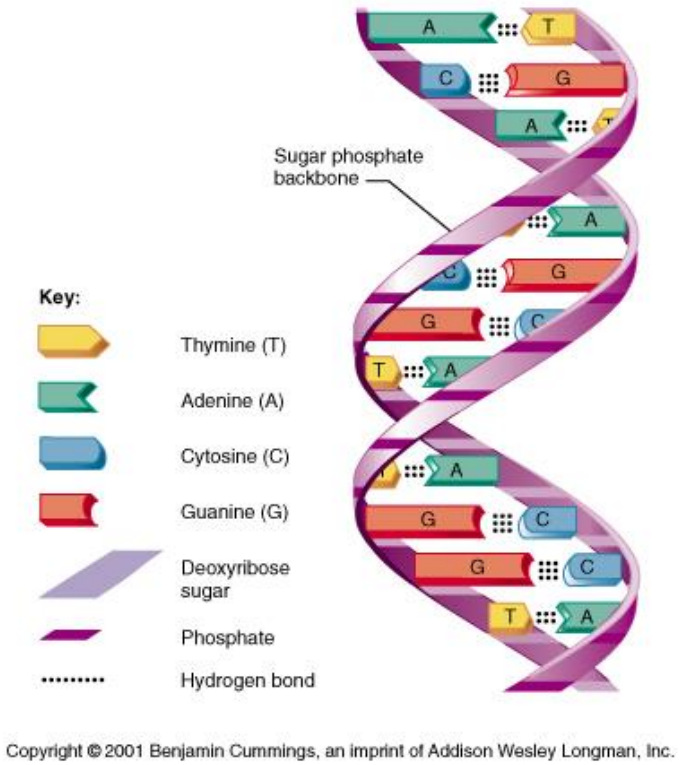

Figure 2.5 Structure of DNA (Babich 2012)

2.4.1 DNA Analysis

DNA analysis is mostly applied on forensic such as forensic DNA analysis that uses of DNA samples in legal proceedings. Just as people can leave fingerprints when they touch a surface, they can leave biological material that contains DNA. When a person's fingerprint matches the latent print found at the scene of a crime, the match provides evidence linking the person to the crime. Similarly, DNA recovered from stains of blood, saliva, or semen or from material such as bone, hair, or skin can be matched to a person's DNA. DNA can even be recovered from fingerprints (Evett & Weir 1998).

Forensic DNA analysis has gone through several stages of development. The first generation of DNA analysis was restriction fragment length polymorphism (RFLP) profiling and it is no longer used by the forensic community, as it requires relatively large amounts of DNA and degraded samples could not be analyzed with accuracy. The second generation of DNA analysis was based on polymerase chain reaction (PCR). However, it is not suitable in the analysis of longer strands of DNA.

The third generation of DNA analysis or the current method of choice is short tandem repeat or STR analysis (Romeika & Yan 2013).

2.4.2 Short Tandem Repeat

Tandem repeat is a pattern that helps to determine an individual's inherited traits. Tandem repeats occur in DNA when a pattern of two or more nucleotides is repeated and the repetitions are directly adjacent to each other (Oki et al. 1999). When two nucleotides are repeated, it is called a dinucleotide repeat (for example: CTCTCTCTCT…), for instance colon cancer most commonly affects such regions. When three nucleotides are repeated, it is called a trinucleotide repeat (for example: CTACTACTACTA…) (Pennisi, 2004). All DNA contain of deoxyribose sugar, phosphate, and the four bases A, T, C and G. What makes everyone different is the order and number of each base pair in their DNA. If one looking at every single base pair in a person's DNA, it would find that no two people have exactly the same sequence. The problem is that the number of base pairs in DNA is so huge that no DNA laboratory in the world can test the entire DNA. It would take too long time to look at the entire DNA (Soram & Khomdram 2010).

Instead of looking at the whole DNA base pairs, workers are looking for certain special areas of DNA. These areas are believed to be parts of the DNA but can be different in different people. There are identical repeats of the same pattern with length of 2 to 6 base pairs of DNA, and they can be repeated anywhere from 1 to 50 times in a row. These repeats are called Short Tandem Repeats (Soram & Khomdram 2010). An example of short tandem repeat is shown in Figure 2.6, in which the sequence is repeated 7, 8, 9, 10 and 11 times.

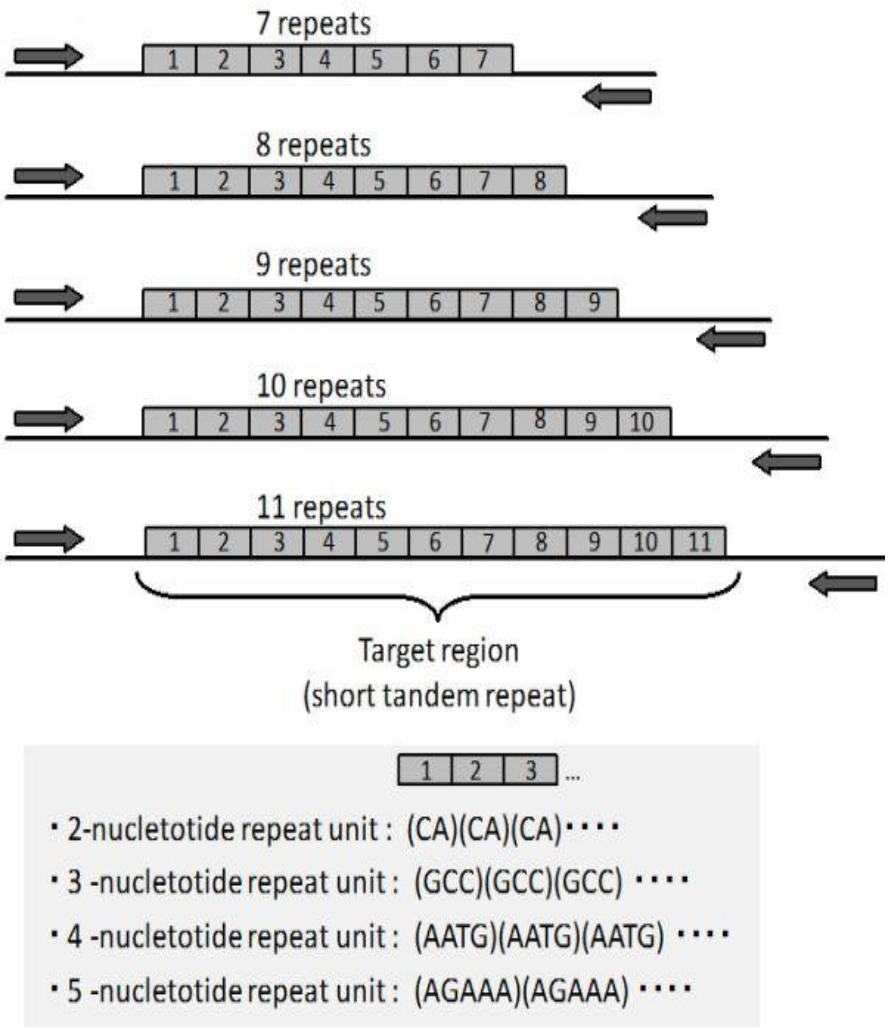

Figure 2.6 Short tandem repeats example (Hashiyada 2011)

2.4.3 Mutation

A gene mutation is a permanent change in the DNA sequence that forms a gene, it makes this sequence differs from the sequence that exist in most people. The genes themselves don't cause disease; genetic disorders are caused by mutations that make a gene function work incorrectly. For example, when someone has “the cystic fibrosis gene”, they are usually indicating to a mutated version of the *CFTR* gene, that lead to the disease (All people, whom they have not the cystic fibrosis; have a version of the *CFTR* gene). Some mutations alter the DNA sequence of the gene but do not change the protein function that established by this gene. The gene DNA sequence can be altered in different ways. Gene mutations have different effects on health; these effects are depending on where they occur and if they alter the proteins function or not

(US.NLM 2018). Some types of mutations are missense mutation, nonsense mutation, insertion, and deletion. Missense mutation is a mutation that changes one DNA base pair that lead to substitution of one amino acid with another in the protein made by a gene, as shown in Figure 2.7.

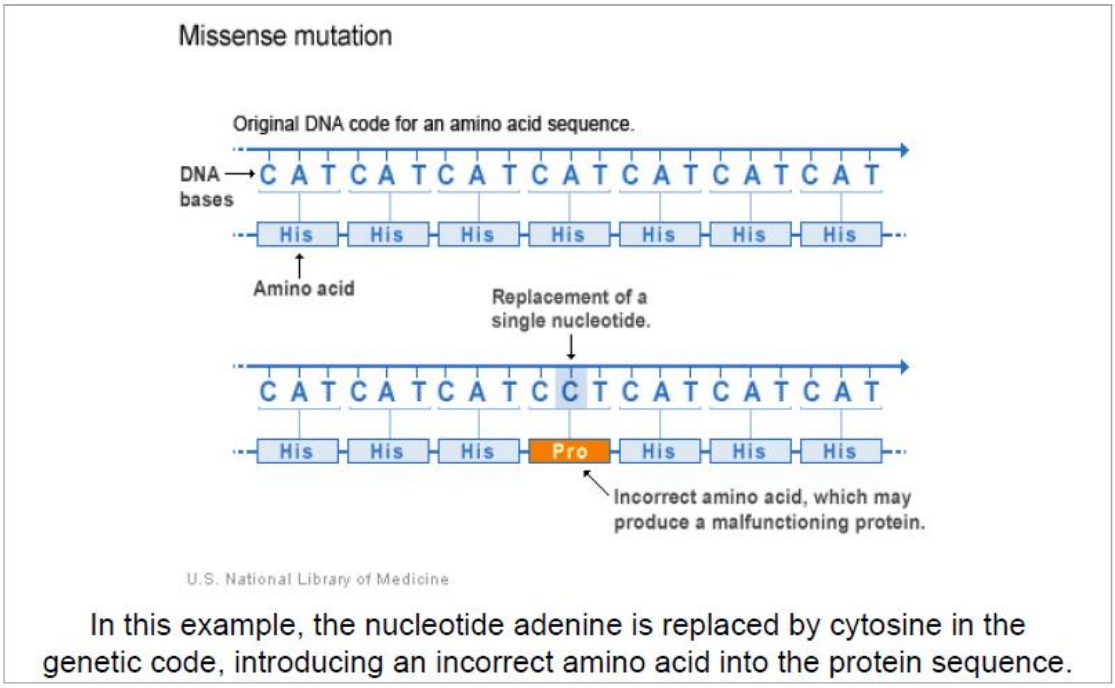

Figure 2.7 Missense mutation example (US.NLM 2018)

The nonsense mutation is also a change in one DNA base pair. Instead of substituting one amino acid for another, however, the altered DNA sequence prematurely signals the cell to stop building a protein (US.NLM 2018). This type of mutation results in a shortened protein that may function improperly or not at all, as shown in Figure 2.8.

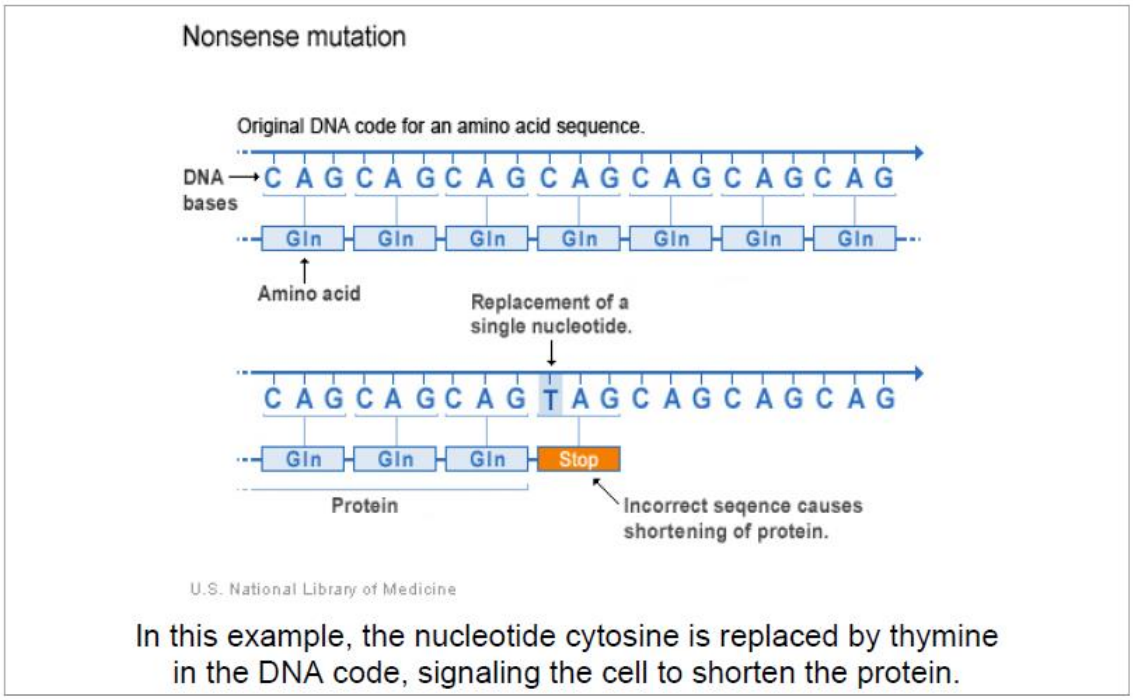

Figure 2.8 Nonsense mutation example (US.NLM 2018)

An insertion mutation changes the number of DNA bases in a gene by adding a piece of DNA. As a result, the protein made by the gene may not function properly, as shown in Figure 2.9.

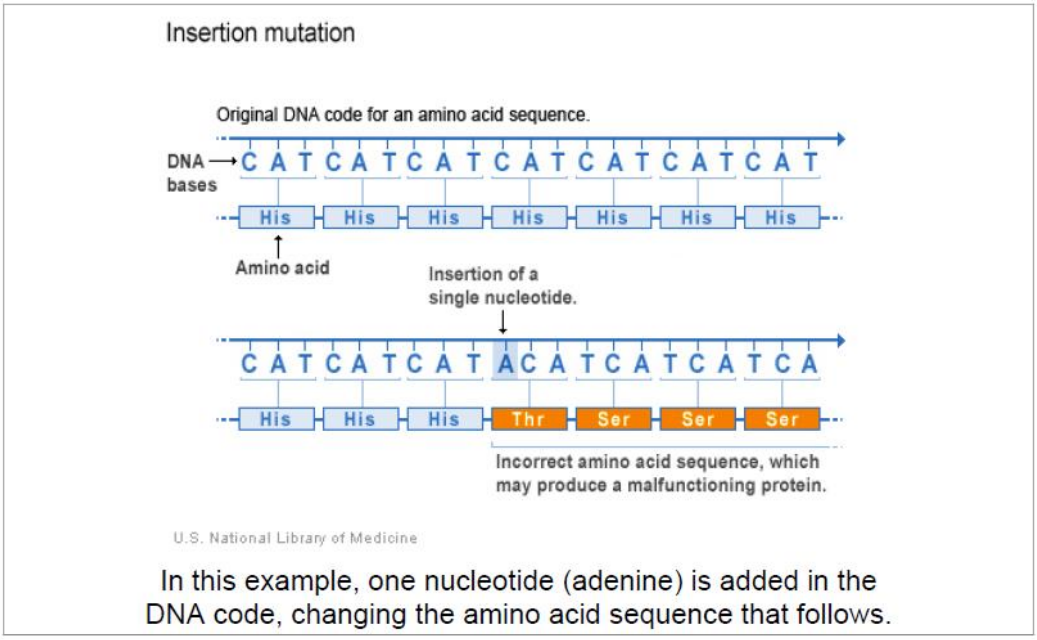

Figure 2.9 Insertion mutation example (US.NLM 2018)

A deletion mutation changes the number of DNA bases by removing a piece of DNA. Small deletions may remove one or a few base pairs within a gene, while larger deletions can remove an entire gene or several neighboring genes. The deleted DNA may alter the function of the resulting protein, as shown in Figure 2.10.

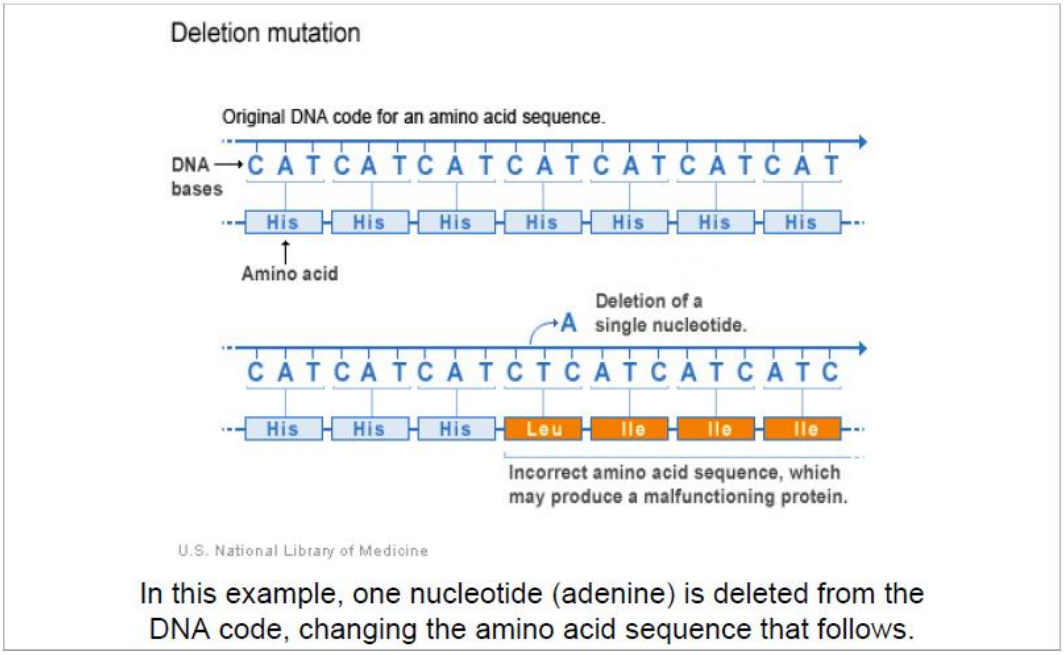

Figure 2.10 Deletion mutation example (US.NLM 2018)

DNA is always subject to mutations that lead to change its code. Mutations can cause losing or deformed proteins that may causes disease. All lives begin with some mutations that inherited from parents. In addition, during lifetime mutations can be acquired, these mutations occurred during cell division when DNA is duplicated. Other mutations are happened when DNA is damaged by environmental factors, such as ultraviolet radiation, chemicals, and viruses (US.NLM 2018). An example of DNA sequence for healthy person and ill person is shown in Figure 2.11; in this example, the mutation happened in the *β-globin* gene, this gene is responsible for the genetic blood disorder β-thalassaemia.

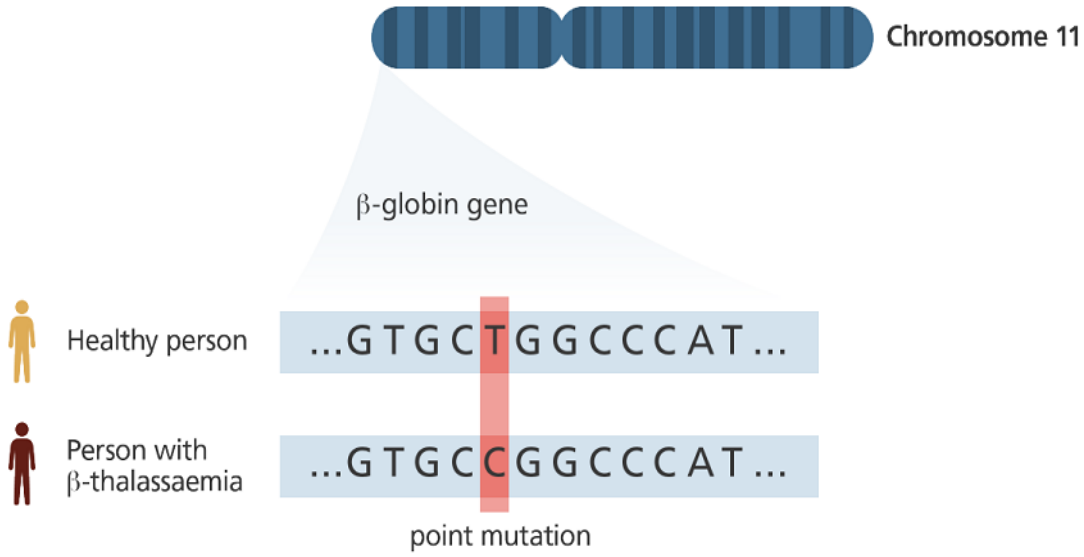

Figure 2.11 An example of healthy person and ill person (Internet-1 2018)

Mutations detection can be identified depending on the results from a genetic analysis using DNA markers such as STR markers flanking or within the gene (Nakamura 2009). Molecular diagnostics provide a way for assessment of the genetic makeup of human; it combines laboratory medicine with molecular genetics to develop DNA/RNA-based analytical methods for monitoring human pathologies. Molecular methods for identification of the disease-causing mutations could be classified as methods for known and methods for unknown mutations. Many different approaches have been used for identifying known mutations. Usually starting with the PCR (Mahdieh & Rabbani 2013), which is a method for amplifying DNA fragment to a large number of fragments in only a few hours (Mullis & Faloona 1987; Eisenstein 1990). PCR; consisting of 25-40 repeated cycles, each cycle has three steps. The first step is denaturation where DNA sample is heated in 92–98 ℃ to separate it into two single strands (DNA melting). The second step is annealing, that performed by decreasing temperature to 50–65 ℃, so the primers are annealed to their targets on single stranded DNAs by hydrogen bonds, and the third step is elongation that includes polymerization of the bases to the primers and copies the strand, this step is done at a temperature of 70–74 ℃. (Mahdieh & Rabbani 2013). Figure 2.12 shows the PCR steps.

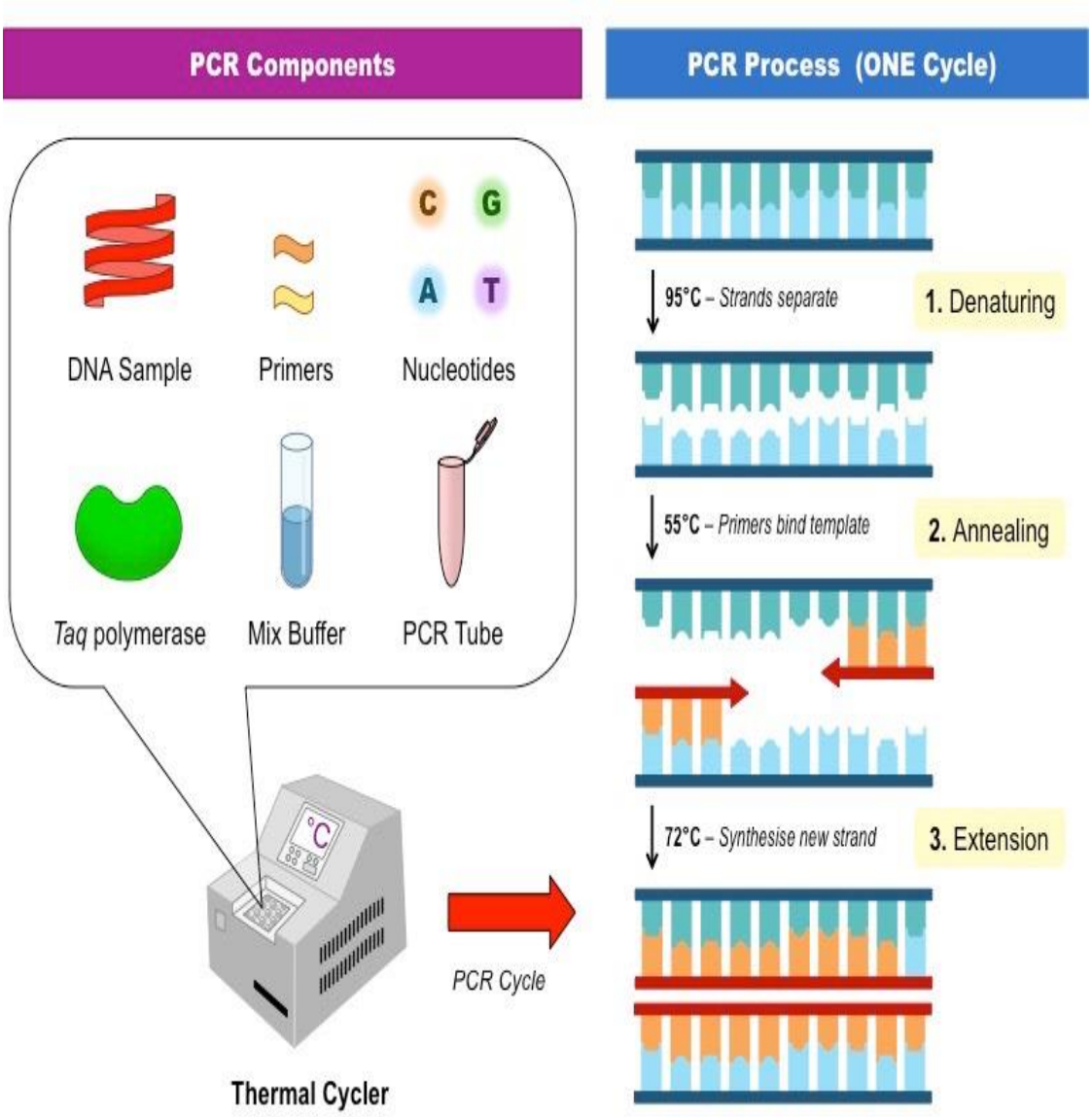

Figure 2.12 PCR steps (Internet-2 2018)

Computers are used for storing and analyzing biological data. A DNA sequence of various diseases is stored in database for easy retrieval and comparison. A specific sequence causes a disease; this disease can be diagnosed by counting the number of patterns matching strings occurred in a gene database. Pattern matching algorithms used to find a specific sequence in the DNA (Rajesh et al. 2010). There are similarities and differences between biological viruses and computer viruses, Tables 2.6 and 2.7 shows these similarities and differences respectively.

Table 2.6 The similarity between computer viruses and biological viruses (Korthof 2015)

No.	Computer Viruses	Biological Viruses
	Similarities	
1	infection of specific targets (.exe or .com files)	infection of specific targets (host cells)
2	attach to .exe or .com files	integrate in DNA
3	spread to other computers	spread to other hosts
4	parasitism: copied by host	copied by host cell
5	one virus per file	no re-infection of same cell
6	initially infected file is functional	initially infected cell is functional
7	user does not immediately notice infection	host organism does not immediately notice infection
8	software can be made immune to infection	not every cell is infected
9	specificity for Operating System	host specificity
10	species (kinds) of viruses	species (families) of viruses
11	degrees of harmfulness	different degrees of virulence
12	difference in susceptibility of computers	difference in susceptibility of individuals and species
13	antivirus software on computer	immune system
14	percentage of computers protected by anti-virus software	percentage of individuals in population immune to virus (vaccinated)
15	PC's came first, viruses later	host organism evolved prior to infecting virus
16	contain information, have length expressed in bytes	contain information, have length expressed in bases
17	source code causes behavior of virus	genotype causes phenotype including behavior
18	virus has small size relative to host software	small genome relative to host genome
	Potential similarities	
1	mutating virus	virus mutates
2	activation of virus depends on date	seasonal activity of virus
3	software version dependent action	age dependent action of virus
4	virus infects new host software	infection of new host species
5	anti-virus software introduced	evolution of immune system
6	anti-virus software comes at a price	immune system is costly for the organism

to be continued…

…continuation

7	arms race virus and anti-virus software	arms race virus and immune system (vaccines development)
8	spread via Trojan horse	spread via vector
9	hidden presence of virus	latency; initial symptom free period
10	polymorphic virus	polymorphic virus
11	virus disables virus scanner	virus attacks immune system
12	Darwinian evolution of mutating viruses	Darwinian evolution of mutating viruses
13	detected by virus signature	detected by virus signature

Table 2.7 The differences between computer viruses and biological viruses (Korthof 2015)

No.	Computer Viruses	Biological Viruses
1	created by humans	created by biological evolution
2	source code known to author of the virus	sequence of new virus not known
3	no 2D or 3D form	always 3D form / structure
4	virtual (digital)	material; based on molecules
5	no auto-immunity	auto-immune diseases
6	useful viruses do not exist	some viruses have useful effects for the host

2.4.4 DNA Computing

Intrusion detection system is used to detect attack when it tries to enter the system by looking for suspicious patterns (that different from user behavior). Similar to DNA, human body checks if there are any changes at specific DNA sequences (that different from the healthy body); if there is a change in DNA sequences that means a disease attack the human body. The DNA contains a huge number of base pairs, while the network traffic contains only 41 features.

DNA can be applied into different computer system techniques such as cryptography, steganography, digital signature, intrusion detection, and others. This done by either using DNA computing or DNA encoding. A DNA computing is a branch of computing system, which uses DNA and molecular biology hardware, instead of the traditional based computer technologies, to solve many problems. Research and development in this area includes theory, experiments, and applications

of DNA computing. Leonard Adleman of the University of Southern California initially developed this field, in 1994. Adleman demonstrated a proof-of-concept use of DNA as a form of computation, which solved the seven-point Hamiltonian path problem (Adleman 1994). Many researchers have used DNA computing in computer system in different fields (Hasudungan & Abu Bakar 2013; Maazallahi et al. 2013; Fasila & Antony 2014; Ibrahim et al. 2014; Hakami et al. 2015; Boruah & Dutta 2016; Sanches & Soma 2016; Mondal & Mandal 2017; Murugan & Thilagavathy 2017; Zhixiang et al. 2017; Kate et al. 2018; Xu et al. 2018).

A DNA computing method is suggested by Hasudungan & Abu Bakar (2013) to solve the distribution center location problem by proposing DNA encoding to represent distribution center location problem (DCLP) data and this operation done by using biochemical reaction. Maazallahi et al. (2013) proposed a new method to solve N-queens problem based on Adleman-Lipton DNA computing model. A new hybrid cryptography method introduced by Fasila & Antony (2014) that depends on RGB colors, the data secured by using encryption algorithm and then used a strong key based on DNA steganography method. Ibrahim et al. (2014) displayed an application of the ant colony system for DNA sequence in DNA computing by using two parameters, these parameters were namely, ΔGo37 and nearest neighbor thermodynamic. Hakami et al. (2015) proposed a possible DNA computing method for big data. This was done by converting data (which can be image, text, or multimedia) to binary format, then converted it to DNA sequences by using DNA encoding, and finally applied DNA decoding method. A DNA algorithm was presented by (Boruah & Dutta 2016) that used to simulate the logic functionality of any Boolean circuit, and it included a hybridization process and also avoids use of process that may achieved an error. Sanches & Soma (2016) presented a DNA computing method based on some biological operations to solve two NP-Hard problems, which are minimization of open stacks problem and matrix bandwidth minimization problem.

Mondal & Mandal (2017) proposed a new system to secure image communication based on a sequence of pseudo random number and DNA encryption. Murugan & Thilagavathy (2017) improved cloud security by proposed a new DNA

computing model based on adding Morse code and Zigzag pattern. A DNA algorithm is proposed by (Zhixiang et al. 2017) to solve integer related programming problem by using double stranded encoding, where the advantages of this method are the simplicity and error-resistant, and feasible. Kate et al. (2018) proposed a novel algorithm for voice encryption by combining three techniques; these are innovative encoding scheme, DNA encryption method, and a permutation technique. Xu et al. (2018) proposed a system to solve the graph vertex coloring problem based on DNA computing model. This was done via divide the subgraphs and reduces the bio-operation time by design a parallel polymerase chain reaction (PCR).

On the other hand, DNA encoding is the process of converting plaintext to DNA sequence. Several researchers have used DNA encoding in computer system in different fields; Al-Ibaisi et al. (2008) built a new DNA encoding tables for IDS by divided the network traffic attributes into: static parameters (that include: flag, service, and protocol attributes) and dynamic parameters (that include: integers, reals, and Booleans attributes). Then, three DNA characters to represented each network traffic attributes values (static parameters and dynamic parameters) are used and put three characters as a header in front of static parameters attributes values. Wang & Zhang (2009) built a DNA encoding for cryptography method, where they firstly used three DNA characters to represent each alphabetic character (uppercase characters that include 26 possible values) and secondly, they used two numbers to represent each DNA character (4 possible values which are A, C, G, and T). Later, Jarold et al. (2013) built a DNA encoding method for message cryptography by used three DNA characters to represent each plaintext character, where the original text (plaintext) can include 40 values (26 characters, 10 digits, and 4 special characters) and these three DNA characters can handle and represent all of these 40 values. After that, Hameed & Rashid (2014) enhanced the DNA encoding method that proposed by Al-Ibaisi et al. (2008) and applied it for IDS, where they used three DNA characters to represent flag, service, and protocol attributes values and put another three DNA characters as a header (that mean each attribute value was represented by six characters). Also, they used two characters to represent the digit attributes values. A new DNA encoding method is built by UbaidurRahman et al. (2015) for text cryptography, where they used four DNA characters to represent each plaintext character. The original text can

include 96 values (52 characters, 10 digit, and 34 special characters) and these four DNA characters can handle and represent all these 96 values. Table 2.8 shows the characteristics to be fulfilled by encoding method (UbaidurRahman et al. 2015). Some research that used DNA encoding in computer system in different fields are shown in Table 2.9.

Table 2.8 Definition of characteristics to be fulfilled by encoding algorithm (UbaidurRahman et al. 2015)

No	Features	Definition
1	Robustness	The encoding method is not breakable.
2	Confidentiality	The DNA encoding method should help to encode the plain text, which insure that this text is not possible to be deciphered.
3	Randomness	A non-order of sequences that achieve more randomness of DNA encoding table generation.
4	Dynamicity	Generates and uses a new encoding table every session between a sender and receiver.
5	Complete Character Set Fulfillment	The DNA encoding table should provide for DNA encoding sequences for the complete character set.
6	Uniqueness	The encoding of plaintext into DNA sequence is unique in every generation of encoding table.

Table 2.9 Literature reviews on DNA encoding for computer security

Author(s)	Year	Technique	DNA Encoding	Total Values	Description
Al-Ibaisi et al.	2008	IDS	Used three DNA characters to represented each network traffic attributes values and put three characters as a header in front of static parameters attributes values	98 values (10 integer, 1 real, 2 Boolean, 11 flag, 3 protocol, and 71 services).	Built IDS based on DNA encoding, first converted network traffic to DNA sequences and second used genetic algorithm to optimize the selection for target solution.
Wang & Zhang	2009	Cryptography	Used three DNA characters to represent each alphabetic character and used two numbers to represent each DNA character.	26 uppercase characters, and also 4 DNA values (A, C, G, and T).	Built a system that can transmit message securely by using RSA algorithm and DNA encoding that used to encrypt the message.
Jarold et al.	2013	Cryptography	Used three DNA characters to represent any value. Encoding example: B=CCA.	40 values (26 characters, 10 digits, and 4 special characters).	Built two methods for encoding messages to DNA, the first one use DNA cryptography alone, and the

to be continued…

… continuation

					second one used DNA for key generation.
Chouhan & Mahajan	2014	Digital Signature	Used one DNA character to represent any two binary digits. Encoding example: 00= A.	4 values (00, 01, 10, and 11).	Establish a system to encrypt and generate digital signature based on DNA Cryptography. This system is done by four steps: first generate the encryption key, then encrypt the plaintext with the key that generated in first step, then use DNA encoding to convert the cipher text to DNA bases, and the final step is generating the signature.
Das & Kar	2014	Image Steganography	Used one DNA character to represent any two binary digits. Encoding example: 00= A.	4 values (00, 01, 10, and 11).	Built a system to hide secret data in an image, and this system will transfer image between two different parties based on two security layers, the first one is DNA sequences and the second one is an image cover.
Ghany et al.	2014	Image Watermarking	Used one DNA character to represent any two binary digits. Encoding example: 01= A.	4 values (00, 01, 10, and 11).	Build a system to hide DNA sequence data into fingerprint image. A discrete wavelet transforms the function, then DNA decoding is used to extract signature from the watermarked image. The generated watermark is invisible
Jain & Bhatnagar	2014	Data Encryption	Used four DNA characters to represent any value. Encoding example: 1=	256 values (numbers from 0-255)	Build an encryption system based on DNA cryptography, this system encrypted

to be continued…

…continuation

			AAAA.		the data (text, image, audio, or video), by converting the data to binary, then convert the binary to its equivalent decimal, and finally encryption based on DNA.
Mokhtar et al.	2014	Image Encryption	Used one DNA character to represent any two binary digits. Encoding example: 00= A	4 values (00, 01, 10, and 11).	Built a new color image cryptography system based on both DNA sequence and Chaos theory. This system firstly converted the pixel value to binary, then converted it to DNA sequence, thereafter, using the logistic map to change pixel value and position, and finally one-time pad method is used to change the pixel value.
Sasikumar & Karthigaikumar	2014	Text Encryption	Used three DNA characters to represent any value. Encoding example: B= CCA.	40 values (26 characters, 10 digits, and 4 special characters).	Established a cryptography system based on both DNA coding and quantum cryptography. First step is used BB84 protocol to encode the message, the next step is key exchange, then applied DNA encryption to generate DNA sequence, finally applied AES algorithm to encrypt the DNA sequence.
Gupta & Jain	2015	Image Encryption	Used one DNA character to represent any two binary digits. Encoding example: 01= A.	4 values (00, 01, 10, and 11).	Built new gray scale image encryption system based on DNA sequence. The system encrypted the image via two

to be continued…

…continuation

					steps, the first step is using DNA bases matrix to get cipher, and the second step is changing the pixel values.
Saranya et al.	2015	Image Encryption	Used one DNA character to represent any two binary digits. Encoding example: 00= A.	4 values (00, 01, 10, and 11).	Built an image encryption system based on both DNA sequence and genetic algorithm, first step is to generate a key, then the logistic map is used to create many number of DNA masks. An encryption process is done in real image to generate many cipher images, and finally the genetic algorithm is applied to find the best DNA mask.
UbaidurRahman et al.	2015	Cryptography	Used four DNA characters to represent any value. Encoding example: a= ACAT.	96 values (52 characters, 10 digit, and 34 special characters).	Build a novel technique for DNA encryption and decryption processes, which contain all attributes that should be distinguishing of a DNA computing. The performance results show the strength of the proposed algorithm.
Zhen	2015	Image Encryption	Used one DNA character to represent any two binary digits. Encoding example: 00= A.	4 values (00, 01, 10, and 11).	Proposed a secure image encryption method by using chaotic system that increase the complexity, then used and mixed eight DNA coding rules in order to enhance the efficiency of image confusion and diffusion.

to be continued…

…continuation

Kumar et al.	2016	Image Encryption	Used one DNA character to represent any two binary digits. Encoding example: 00= A.	4 values (00, 01, 10, and 11).	Proposed new image security algorithm based on elliptic curve cryptography and DNA encoding. This done by encode the RGB image using DNA encoding, and then make asymmetric encryption based on Elliptic Curve Diffie–Hellman Encryption.
Chai et al.	2017	Image Encryption	Used one DNA character to represent any two binary digits. Encoding example: 00= A.	4 values (00, 01, 10, and 11).	Proposed image encryption algorithm by using chaotic system and DNA sequence operations. Encode plain image into DNA matrix, then performed permutation scheme, and then applied image diffusion at DNA level is applied.
Mumthas & Lijiya	2017	Video Steganography	Used one DNA character to represent any two binary digits. And then used one alphabet character to represent three random DNA characters.	4 values (00, 01, 10, and 11).	Presented a steganography system based on RSA algorithm, DNA encryption, and Huffman encoding by using video to hide the messages. The use of random DNA encryption has improved the system security.
Chen et al.	2018	Image Encryption	Used one DNA character to represent any two binary digits. Encoding example: 00= A	4 values (00, 01, 10, and 11).	Proposed an image encryption system based on random DNA encoding by converting image pixels to DNA sequence then applied permutation procedure for further encryption

to be continued…

…continuation

Junxin et al.	2018	Image Encryption	Used one DNA character to represent any two binary digits. Encoding example: 00= A.	4 values (00, 01, 10, and 11).	Presented a new image encryption system based on permutation– diffusion and DNA encoding. The performance results showed that cryptosystem is secure against plaintext attack.
Rehman et al.	2018	Image Encryption	Used one DNA character to represent any two binary digits. Encoding example: 00= A	4 values (00, 01, 10, and 11).	Built an image encryption system based on DNA encoding with addition of substitution phase to get higher efficiency.
Zhang et al.	2018	Image Encryption	Used one DNA character to represent any two binary digits. Encoding example: 11= A	4 values (00, 01, 10, and 11).	Proposed image encryption method based on Feistel network and DNA encoding technology, the first step is calculating hash value by using SHA-3 algorithm then generate the Hill cipher matrix, the second step is using DNA sequence operation for Feistel network, and the last step is making more diffusion.

The method that used to extract the STR is called biological sequences analysis. The sequence analysis can be used to find the similarity between sequences, to identify sequence features such as sites, and identify sequence differences such as mutations (Gupta et al. 2014).

2.4.5 Pattern Discovery Method

Biological sequences analysis is used to locate the similarities between different sequences, where several computational algorithms exist to discover the similarity of