Network Intrusion Detection by hashed pattern matching using DNA bases encoding

Thaer AL Ibaisi, Muhammad Kazim, Mustafa Kaiiali, Stefan Kuhn Contact: thaer.ibaisi@my365.dmu.ac.uk

DE MONTFORT UNIVERSITY LEICESTER

UNIVERSITY OF TARTU 1632

Introduction

- A Network Intrusion Detection System (*NIDS*) is a host-based software installed to detect attacks on computer network components.
- *DNA encoding* is the process of transforming plain text into DNA or Amino Acids sequences.
- Efficient pattern matching can then be performed using sequence alignment algorithms & Protein Secondary Structure similarities.

Objectives

- Propose a design for Realtime identification of network attacks using Amino Acids encoding.
- Build Protein signatures database for known attacks.
- Utilize sequence alignment methods & hash pattern matching to search suspicious network transactions signature in attacks database with minimal matching time & acceptable false positive ratios.
- Utilize Protein secondary structure similarities and functional domain areas to group similar attacks.

Work Methodology

I- Real Protein Sequence Analysis Phase:

1. Download Protein Data Bank Fasta[7] file that has viral sequences.
2. Apply protein analysis on each sequence and generate secondary structure attributes.
3. Apply data analysis procedure to extract Amino acids encoding rules to have Mapping Catalogue.

II- Network Transactions Encoding & Modeling Phase:

1. Collect database of known and UpToDate network attacks.
2. Apply dimensionality reduction on network attributes to select most effective set similar to protein secondary structure attributes count.
3. Use Amino acids mapping catalogue from ***phase I*** to transform network transactions (Attacks Database) into biological protein sequences.
4. Apply protein analysis on each sequence and generate secondary structure attributes.
5. Combine both data sources together to have labeled protein structure attributes.
6. Build machine learning model to train it on protein secondary structure attributes to learn the patterns of viral proteins.
7. For unseen new Network transactions, it will be encoded into proteins using mapping catalogue, generate structure attributes, then apply same model prediction to label it as attack or benign.

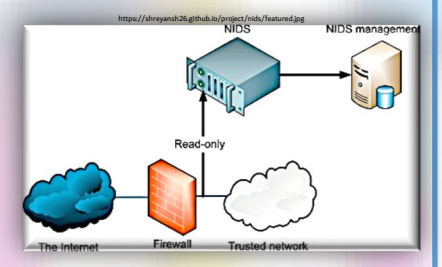
Literature Review

- **Network Intrusion Detection**

The purpose from NIDS is to detect network breaches. It needs to be hosted in the network and check real-time transactions for suspicious activities[1]. For that, NIDS can come in two types:

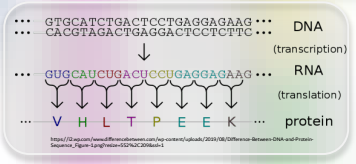
a) Misuse (signature-based) NIDS which uses pattern matching method against pre-defined database of attack DNA signatures.

b) Anomaly NIDS which detects attacks through observing abnormal user activities/patterns [2]. Our research focuses on hybrid type which is composed of both above types.



- **DNA Encoding**

Bioinformatics is science of utilizing bio-inspired computational operations to translate biological information[3]. DNA(*deoxyribonucleic acid*) is series of chemical bases (*A:adenine, C:cytosine, G:guanine, and T:thymine*) that encode genetic information in living cells, which is transformed into mRNA(*Messenger ribonucleic acid*) that by its turn translated into Amino Acids to formulate Proteins. Using DNA to encode digital information took its interest from the ability of little amount of DNA substance to store hundreds of exabytes of digital data in stable manner[4].



Many mappings proposal have been presented by researchers to encode Network transactions and their attributes into DNA nucleotides like:

a) Fixed-length DNA sequence for each command[5].

b) Convert Network transaction into Hexadecimal format then to map each hex digit to fixed-length DNA sequence[6]. Having DNA sequence makes it ready to be translated into corresponding Amino acids sequence then to shape proteins.

Exact pattern matching can omit catching DNA signatures similarities when having few mismatches that may don't affect the cell type classification.

References

[1] D.E. Denning, "An Intrusion Detection Model", IEEE Trans. Software Eng.,. 13 (1987), pp. 222–232.

[2] N. Hubballi, V. Suryanarayanan "False alarm minimization techniques in signature-based intrusion detection systems: a survey", Comut. Commun., 49 (2014), pp. 1-17, 10.1016/j.comcom.2014.04.012

[3] D. Benton, Bioinformatics—principles and potential of a new multidisciplinary tool", Trends Biotech. 14 (1996), pp.261–312.

[4] R. R. Garafutdinov, D. A. Chemeris, A. R. Sakhabutdinova, O. Y. Kiryanova, C. I. Mikhaylenko, and A. V. Chemeris, "Encoding of non-biological information for its long-term storage in DNA," Biosystems, 215–216 (2022), p. 104664, doi: 10.1016/j.biosystems.2022.104664.

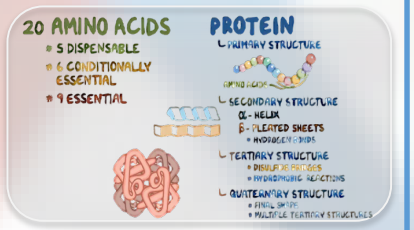
[5] H. Cho, S. Lim, V. Belenko, M. Kalinin, D. Zegzhda, and E. Nuralieva, "Application and improvement of sequence alignment algorithms for intrusion detection in the Internet of Things", in 2020 IEEE Conference on Industrial Cyberphysical Systems (ICPS), pp. 93-97, doi: 10.1109/ICPS48405.2020.9274752.

[6] K. H. Alnafisah, "An Algorithmic Solution for Storing Big Data on the DNA Sequence", in 2020 IEEE 6th World Forum on Internet of Things (WF-IoT), 2020, pp. 1–4. doi: 10.1109/WF-IoT48130.2020.9221056.

[7] https://www.rcsb.org/downloads/fasta

[8] https://biopython.org/

- **Proteins Structures Analysis**
  - Protein structure analysis is an important field of research that aims to understand the three-dimensional structure of proteins and their functions.
  - To analyze protein structure, we used computational tools and algorithms to predict the secondary structure and other important features of a protein sequence.
  - We then calculated the Protein structure attributes like molecular weight, aromaticity, and alpha-helix for each protein using Biopython library[8].



- **Neuro Network Intrusion Detection System**
  - Main approach is to design a machine learning model using Neural Network to train it on Protein structure attributes to learn the patterns of viral signatures.
  - This model will be used to predict the Attack Network Transactions based on protein structural attributes of their encoded Amino Acids.
- **Contribution & Innovation**
  - Our research goal is to highlight & prove the relationship between viral protein patterns in nature and Computer Network Attacks.

Future Perspective

- Evolutionary genomic analysis is a major subject in the research area of bioinformatics, which it can be utilized in intrusion detection theory.
- Having protein signatures for network attacks with structural information provided, we can predict new attacks whose sequences have similar structure with known ones.
- Machine learning can be used to develop new protein structures. In a similar way, new entries for the attacks database could be created to keep it up-to-date.