2010 Third International Conference on Knowledge Discovery and Data Mining

Classifying Ethernet Data Packets Based on Raw Bit Patterns

William D. Kenworthy
School of Information Technology
Murdoch University
Perth, Western Australia
W.Kenworthy@murdoch.edu.au

***Abstract*--Currently most operations on network data packets are controlled by the applicable protocols such as TCP/IP. However, there is scope to examine and classify the data without resorting to processing through a protocol stack. To do this, use can be made of the complex and sophisticated algorithms developed for the analysis of biological and genomics data. This makes use of similarities in the way information is stored in biological structures and network data traffic. It can be shown that network data flows have many of the same structural characteristics as biological DNA - areas of conservation (an area of data that has the same composition as an area in another packet of data will often have similar functionality), "motifs" with particular functions and the equivalent of "junk DNA" - areas where seemingly random changes occur. This paper looks at the novel application of algorithms designed to process DNA data to analyse and classify Ethernet network data packets based on the patterns discernible in the data rather than the more traditional method of matching fixed fields within the data based on protocol specifications. We are able to show that these algorithms are able to successfully and accurately classify packets of data into groups whose members have similar characteristics based on actual content rather than meta-data. This provides a unique and useful method of grouping and classifying packets that could be of use in diverse applications such as IDS systems, and the search for, and identification of specific types of data.**

Keywords-Communication System Security, Data Communication, Internetworking, search methods

I. Introduction

The work described by this paper shows that it is possible to classify network data packets according to relatedness based on the inherent structure of the data carried by the packet using tools developed for analysing biological data. Computer users often give biological and medical terms to physical and software based parts of their computing equipment. Examples in common usage include "virus" (a malicious computer program that can cause damage to data [1]) and "worms" (a self-replicating computer program, similar to a computer virus, which often use a computer network to propagate themselves [2]) This action implies that there is some form of similarity between biological and computational systems that users easily recognize - the common perception being that this is only a superficial relationship. However it can be shown that there are many characteristics in Biology and Computing that have an obvious and real similarity. The relevant example here is the concept of a genome, with the information being stored within it having many parallels (though inevitably differences as well) with a stream of computer network data. Bioinformatics is the scientific discipline involved in analysing and processing biological data using computational techniques – often called computational biology.

A genome (in bioinformatics terms) is a linear sequence of characters (or symbols) representing the chemical molecules present. The symbol table (or “alphabet”) used consists of 4 basic symbols, represented by the letters A, C, G and T, each representing a single molecule. The various components that make up the informational parts of a genome are encoded as patterns in the data using these symbols. Networking data has a direct parallel where the information in a stream of data is encoded as patterns of ones and zeros (binary data). In this paper, we show that this similarity in structure between how information is represented in genomic data, and in network data allows the use of some of the very sophisticated algorithms developed for computational biology to be used to identify and classify useful patterns within Ethernet network data.

II. Previous Work

The methods that are investigated here can be broadly grouped into the area that Bioinformatics practitioners term the "alignment" of two or more sequences of symbols containing common (between the sequences) patterns within the data. Related to alignment is the function of “search which is the process of examining a database of genomic sequences for its nearest match in biological terms to a search term (as a sequence of symbols). An important point to note here is that the matches are statistically ranked based on a metric representing the “relatedness”. A sequence in this context is a variable length set of characters from a symbol set (sometimes called an “alphabet”). The DNA or nucleic alphabet consists of 4 letters - A,C,G and T, each letter representing a particular biological molecule. Software programs such as the BLAST (Basic Local Alignment Search Tool [1]) can do searches and comparisons in a relatively fast and efficient manner. The matches are then ranked in terms of quality and reliability using standardized statistical methods. In biology, molecules can be deleted, added or substituted from DNA based on observable likely-hood. This makes searching difficult as matches then

978-0-7695-3923-2/10 $26.00 © 2010 IEEE
DOI 10.1109/WKDD.2010.83

become “fuzzy” with a metric of “distance” based on changes required to change one sequence of symbols into another required to enable relationships to be viewed.

Alignment is the process of comparing two or more sequences in order to place each of the components in all the sequences into a correct relationship with each other. "Search" and "alignment" are related in that similar algorithms are usually used for each task - an alignment being the aligned comparison of one sequence of symbols with another sequence of symbols in order to obtain their true relationship and quantify the similarities and differences. Search implies an alignment of one "query" sequence with every other sequence of symbols in a set ("database") of sequences in order to rank and quantify the relationship between the query, and all other sequences in the database. A search then returns the best, or closest matches in statistical terms, ranked in order of importance using the selected metric.

One of the key concepts in the analysis of genomic data is that of "conservation" [3]. This is where biological genes with a similar functionality or purpose have motifs (the meaning of motif in this context is "short, well defined patterns in the data") in common. Further, these "motifs" are stable in that any changes to their pattern or make-up can change the functionality of the genome in which they form a part which can often have a deleterious effect on the organisms "fitness to survive" [4]. Parts of individual network data packets may be viewed as an organizational analogue to a feature in a genome. On a genome, sequences of molecules form patterns which are therefore features on the genome. By "organizational analogue", we are referring to the fact that a packet of network data has sections that have a defined function and is a recognizable pattern in the flow of data.

To illustrate the problem, network packet headers such as those used by the TCP/IP protocol [5] have a particular function allocated to each bit, or combination of bits in the header. An example is the 32 bit value defining the destination address in an IP (Internet Protocol [5]) packet. Any change to the bits in this area will change the destination address of the packet (ignoring the fact that the packet will be discarded as defective by the network protocol stack because of a CRC mismatch) - this is equivalent to a “mutation” [4] in a genome, and causes a noticeable change in the packets functionality. Similarly, there are areas in some genomes termed "Junk DNA" [4] - that is DNA which seems to have no purpose, or organizational structure, and often can be very unstable (many changes or "mutations" occur over short periods of time). This biological term has a direct analogue in the payload or data area of an Ethernet packet which changes continually from packet to packet.

III. Methodology

Standard bioinformatics programs and networking utilities were used for this project along with the addition of a simple Perl script(s) to pre-format the networking data to make it compatible with the most common file storage formats used in bioinformatics.

The key difference between network data and genomic data are the symbol sets or alphabets used. DNA uses the four character symbol set "A", "C", "G" and "T" [3] representing four different molecules (or "states"), while network data packets are essentially binary (two "state") data, "0" and "1". The simple conversion used here was to map the "0" to "A", and "1" to "T" in order to use existing, unmodified bioinformatics software. More complex mappings may work better - but the above appears to work well in initial testing. More detailed investigation will require native data formats rather than the proof of concept conversions used here. The output of the above process is a "FASTA" [6] formatted file containing the "A"’s and "T’s", each representing the corresponding binary ones and zeros of the data packet, along with a FASTA format header with details to identify the data packets.

The packets were captured "live" using the "tcpdump" utility to create a "raw" packet capture file (known as a "packet dump"). The capture file was then offline processed using the "wireshark" [7] network data packet capture and protocol dissector software application, the processed output of which is in a annotated, hexadecimal format. This was converted to the genomic FASTA format by a custom Perl script - the output of which is a standard file ready for processing by standard, unmodified bioinformatics programs.

The next stage in this process is to "align" the data packets in order to juxtapose similar motifs across each of the packets of data. The program used for this is "ClustalW" [8]. ClustalW is a heuristic "global" alignment program used for Multiple Sequence Alignments or "MSA". ClustalW’s global approach is particularly suited to this purpose as it returns the complete sequence (areas both in alignment, and those areas outside the detected alignment, not just the areas of local similarity as returned by "local alignment" algorithms. Internally, ClustalW uses many sophisticated techniques to get the most accurate alignment. There are also many options controllable by commandline switches that are not used here, but offer avenues to investigate the impact of “substitution” [4] and other advanced techniques in the future. The formatted example of this seen at Annex B clearly shows the areas where motif’s are shared between the data packets.

Lastly, a phylogram of the packet relationships is generated using the "PHYLIP" [9] group of programs. The "dnadist" [9] program (using the Kimura 2 parameter method) was used to generate the computed distances between each packet of data, and the "neighbor" [9]

program (using the neighbor-joining method) parameters to calculate the relationships between the data packets with the final outputs being a set of metrics that characterize the differences and similarities between each of the data packets, which enables the generation of a graphical "plot". The plot shows each packet in a spatial relationship with every other packet, enabling groupings to be quickly and unambiguously assessed. Because of this similarity in structure, bioinformatics algorithms that target conserved areas (search and alignment algorithms) can be used to analyse and classify network data.

IV. Results

The results shown here are based on a small (14 packet) capture of packets on a network. The two figures involved display the results in a form suitable for examination by the human eye, but the ultimate aim is to use the data and statistics from the algorithms to form an automatic detection system for specific packet types in an IDS. In Fig. 1 which shows part of a "multiple sequence alignment" (MSA) using the "ClustalW" program [8] which displays 14 packets of network data showing the areas of conservation (characters in common or identity across two or more packets of data).

The numbers on the left of the list are an index number allocated to each packet of data based on order of collection. The index entries are grouped and sub-grouped according to similarity. Index 12 has no identity in the area displayed - those above index 12 form one group, while those below index 12 form another group. Packets 11 and 14 are different types of SAP broadcasts while packet 13 is a Cisco broadcast hence being grouped together based on being broadcasts. Packet type and content was verified using the "WireShark" software tool.

```
2  01000000000000000010101101100111011001011101000000----000000
8  01000000000000000010111101100111011001101101000000----000000
4  01000000000000000010101110100111011001100001000000----000000
6  01000000000000000010101110100111011001101001000000----000000
1  01000000000000000010010001011111011010111101000000----000000
7  01000000000000000010010011011111011011011101000000----000000
3  01000000000000000010001111011111011011000001000000----000000
5  01000000000000000010001111011111011011010001000000----000000
9  01000000000000000010010001011111011011000101000000----000000
10 01000000000000000011010000100111011001110001000000----000000
12 ------------------------------------------------------------
11 11111111111111111111111111111111111111111100000000-------010
14 11111111111111111111111111111111111111111000000001110000010
13 11111111111111111111111111111111111111111000000000001001010
```

Figure 1. Portion of an MSA of 14 Ethernet Frames

Fig. 2 is a "plot" of each packet of data showing the "distance" between each packet of data as a function of the number of observed differences between each packets. The

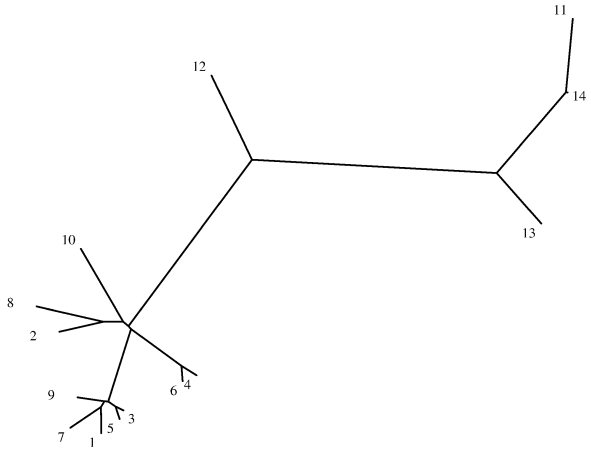

Figure 2.Plot of network packet relationships based on calculated differences

"distance metric is based on the number of changes between any two points on the graph. The shape is "bifurcating" and arbitrary being the result of the layout program (in this case using the "drawtree" package from the PHYLIP package [9]). Visually, packets with similar characteristics cluster together on a branch - the tightness of the clustering is a function of the number of binary differences in position and value between the packets. If protocol headers such as TCP/IP are included along with the payload data, this will cause packets to/from similar addresses to cluster closer together. Even though this information is only a few percent of the full packet size, its constant content and placement means it has a "high" similarity causing packets to cluster with other packets to/from the same hosts. If pre-processing is used to remove header information, clustering can show characteristics that is purely due to payload structure.

V. Conclusions

This paper demonstrates a novel and alternative way of classifying packets of data in a way that is able to show the "relatedness" of packets as a function of the number of differences between them via their data structure/characteristics rather than the yes/no results of applying normal protocols. This enables the classification/examination of data that may be non-standard, corrupted or unrecognisable as it is incomplete by looking at structures within the data itself.

VI. References

[1] Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W. & Lipman, D.J., "Basic local alignment search tool.", Journal of Molecular Biology, vol 215, pp. 403-410, 1990

[2] Phoha, Vir V., "Internet Security Dictionary", pp. 139, 141, Springer-Verlag, 2002.

[3] Mount, David W., "Bioinformatics: Sequence and Genome Analysis", pp. 70-71, Cold Spring Harbor Laboratory Press, 2004.

[4] NCBI, "NCBI Handbook", Available at http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=hand, 2003.

[5] Information Sciences Institute, "RFC793: TRANSMISSION CONTROL PROTOCOL: PROTOCOL SPECIFICATION", 1981

[6] NCBI, "fasta Format description", Available at http://www.ncbi.nlm.nih.gov/blast/fasta.shtm, .

[7] Combes, G., "WireShark", Available at http://www.wireshark.org/, 2006.

[8] Ramu Chenna , Hideaki Sugawara , Tadashi Koike , Rodrigo Lopez , Toby J. Gibson , Desmond G. Higgins , and Julie D. Thompson, "Multiple sequence alignment with the Clustal series of programs.", Nucleic acids research, vol 31, pp. 3497-3500, 2003

[9] Felsenstein, J., "PHYLIP - Phylogeny Inference Package (Version 3.2).", Cladistics, vol 5, pp. 164-166, 1989