Inter and Intra Organism Patterns in Proteomic Sequences

Ankit Kr. Pathak#1, Pragya Jaiswal#2, Prashant Sinha#3, Simmi#4

#Cluster Innovation Centre, University of Delhi  
Delhi, India

[1ankitkrpathak@yahoo.com](mailto:1ankitkrpathak@yahoo.com), 2[pragya.jswl@gmail.com](mailto:pragya.jswl@gmail.com),

3[prashant@ducic.ac.in](mailto:prashant@ducic.ac.in), 4[simmimourya@ymail.com](mailto:simmimourya@ymail.com)

Abstract— This document gives formatting instructions for authors preparing papers for publication in the Proceedings of an IEEE conference. The authors must follow the instructions given in the document for the papers to be published. You can use this document as both an instruction set and as a template into which you can type your own text.

1. Introduction

This document is a template. An electronic copy can be downloaded from the conference website. For questions on paper guidelines, please contact the conference publications committee as indicated on the conference website. Information about final paper submission is available from the conference website.

1. Page Layout

An easy way to comply with the conference paper formatting requirements is to use this document as a template and simply type your text into it.

1. Page Layout

Your paper must use a page size corresponding to A4 which is 210mm (8.27") wide and 297mm (11.69") long. The margins must be set as follows:

* Top = 19mm (0.75")
* Bottom = 43mm (1.69")
* Left = Right = 14.32mm (0.56")

Your paper must be in two column format with a space of 4.22mm (0.17") between columns.

1. Methodology

We employed different methodologies for the sub-hypothesis that were defined in previous section.

1. (Hypothesis 1) Reestablishment of Evolutionary Taxonomy

We obtained a sample of 59 organism’s (Eukaryotes) proteome sequences from Ensembl Genome Browser, 17 viral[[1]](#footnote-1) proteome sequences and 18[[2]](#footnote-2) bacterial proteome sequences from Uniprot Catalog. The resulting sequence was pre-processed and the amino acid counts were calculated. We used several statistical tools to establish our hypothesis.

1. Pre-processing:

The proteome sequences for the sample were processed to find out the respective count of each of the 20 (and undetermined 21st) amino acids. We define a count vector that stores this respective count in a 21 dimensional vector.

Further, we scaled the vector to get the relative ratios of each amino acid. This was done due to the fact that the counts of individual amino acids change drastically between different organisms (on the basis of their biological complexity). We define the scaled count vector as:

We use vector for all the statistics under this hypothesis.

1. Plotting the Vectors:

Due to the large dimension (21) of the vector we used the parallel plot technique to obtain the line charts. This was done for a visualization of the relatively high dimensional vectors. The parallel axes were set to a common scale.

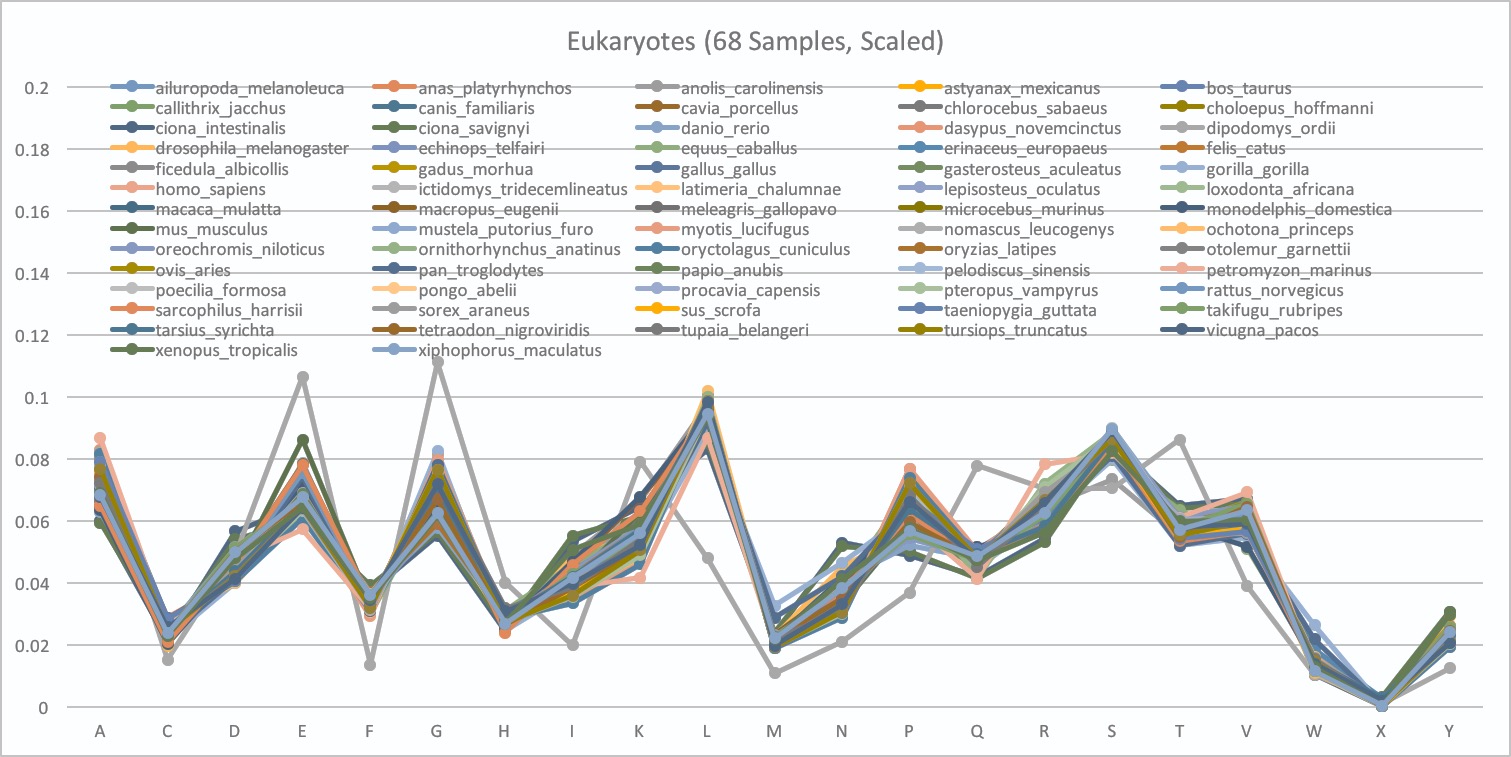
1. Pearson product-moment correlation coefficient:

To quantize the observation in the plots obtained we calculated four sets of correlation coefficients. The sets are:

* Correlation of Eukaryotic Proteome against itself,
* Correlation of Bacterial Proteome against itself,
* Correlation of Viral Proteome against itself, and
* Correlation of Eukaryotic Proteome against Viral and Bacterial Proteome.

Further, the mean of the correlation coefficients obtained was calculated. A higher value of this mean would indicate a greater closeness between the vectors. A lower value of the correlation coefficient, on the other hand, would indicate the inherently distinct behaviour of the data.

1. k-Mean Clustering:



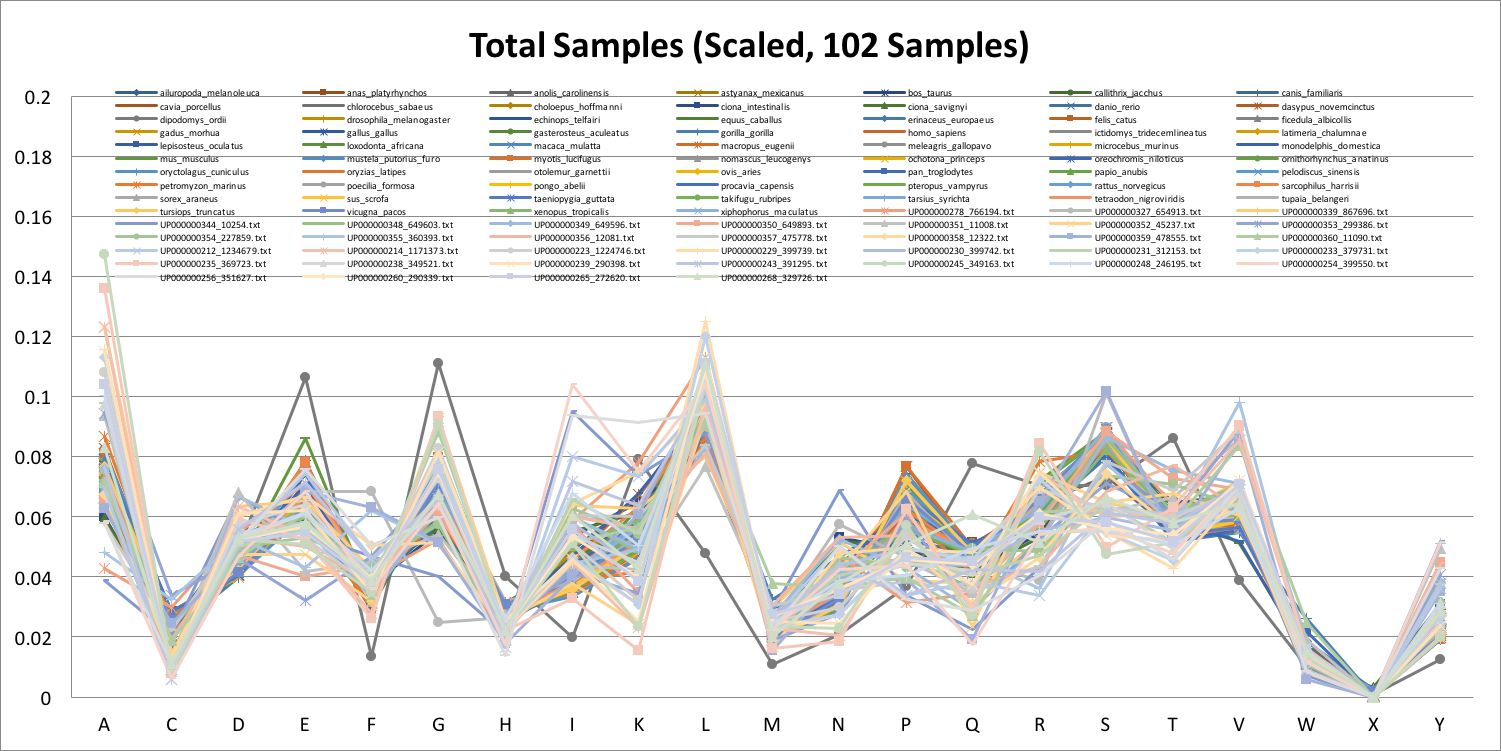


Figure 1Table s

To further strengthen the closeness between the vectors, we ran a k-mean partitioning with all vector samples. A k-Mean object was trained using the concatenation of Eukaryotic Proteomes. By re-running the samples on the k-mean fitted object and counting the number of respective partitions obtained, we were able to judge the clustering property of the vectors. Lower number of partitions would indicate greater cluster formation.

1. Principle Component Analysis:

We calculated the first two Principal Components of each vector classes separately. Different components for different classes of vectors indicated the data’s dissimilarity or similarity based on the coordinates they clustered at.

## (Hypothesis 2) Bacterial and Viral targets

Viruses need a host to survive in. The above sample of 429 viruses, obtained for our previous hypothesis, was further tagged according to their hosts. The two subsamples considered were the viruses that attack Bacteria and those that attack *Homo sapiens* (Humans). Alongside, the proteome sequence of *Homo sapiens* and bacteria was taken for comparison. The sample length of bacteria was 250, obtained after random sampling from 692 bacteria.

The pre-processing of the data follows from the previous section.

1. Plot:

The transformed data of a sample of 250 bacteria, 49 viruses and *Homo sapiens*, 117 viruses was plotted and the mean of each was calculated for an initial visualization of a general trend in each of the samples. Parallel plots were obtained with 20 amino acids (and the undetermined 21st) as X-axis and the ratio of the amino acid in the organism as the Y-axis. For comparison, the line plot of the mean of bacteria was plotted alongside the mean of the viruses that attack bacteria. A similar operation was performed for *Homo sapiens* and the viruses that attack *Homo sapiens*.

### k-Mean Clustering:

To quantitatively analyse the closeness or similarities in the plots, we applied the k-mean clustering algorithm. Three different k-Mean objects were fitted with the concatenation of the following transformed vectors.

* Bacteria and the viruses that attack bacteria
* Homo sapiens and the viruses that attack homo sapiens
* Viruses that attack Bacteria and those that attack Homo sapiens.

Predicting the cluster of the respective samples with their trained k-mean object and then counting the number of respective partitions obtained would determine the similarities. A partition indicates the similarities in that data. This was done for all the three k-mean objects above.

In this case, the favourable result is the attainment of no clean clusters from the data that would imply that there is enough similarity in the data for the algorithm to completely tag it in a particular set.

## (Hypothesis 3)

Past studies have established the generality of this central principle of biochemistry that sequence of amino acids in protein specifies conformation. The dependence of conformation on sequence is significant because of the connection between conformation and function of a protein.

It is seen that residues such as alanine, glutamate, and leucine tend to be present in a helices, whereas valine and isoleucine tend to be present in ß strands. Glycine, asparagine, and proline have a propensity for being in turns.

The proteome data of Mycobacterium Tuberculosis was taken in the FASTA format from UniProt database of proteomes. All the possible amino acids sequences of length five and their respective number of occurrences in the proteome were processed. The sequences were ordered in the decreasing order of their frequencies.

The resulting sequences with highest frequencies were analysed and then compared to known motifs of *Mycobacterium Tuberculosis*.

# Results and Discussion

The following sections summarize the results obtained for the three hypothesis.

1. (Hypothesis 1) Reestablishment of Evolutionary Taxonomy
2. Plots

### Pearson product-moment correlation coefficient:

The value of correlation coefficient is higher for Eukaryotic proteome than the Bacterial and Viral proteome. The following table summarises the result of PPMCC test. An important result here is that the values obtained here conform to the arguments put up from the plots.

TABLE I  
Correlation Coefficients and Interpretation

|  |  |  |  |
| --- | --- | --- | --- |
| Candidate | Value | Interpretation | Comments |
| Eukaryotic Proteome against itself | 0.9737 | High correlation (closeness) between the vectors. (Rank 1) | Explains the closeness as obtained in the data. |
| Bacterial Proteome against itself | 0.8813 | Less correlation between the vectors. (Rank 2) | Explains observable irregularities in the plotted vectors. |
| Viral Proteome against itself | 0.8283 | Lesser correlation between the vectors. (Rank 4) | Explains the high irregularities in the vectors. |
| Eukaryotic Proteome against Viral and Bacterial Proteome | 0.8802 | Less correlation between the vectors. (Rank 3) | Explains the irregularities. Strengthens the hypothesis that the vectors are deviating. |

### k-Mean Clustering:

This result indicates that the samples cluster in three distinct partitions.

TABLE II  
Statistics from k-Mean Clustering

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Sample | Clustered | Total | Ratio | Interpretations |
| Eukaryotic Proteome | 59 | 59 | 100% | All the sampled Eukaryotic Proteomes clustered in same partition. (Rank 1) |
| Bacterial Proteome | 15 | 17 | 88.2% | Most of the samples clustered in same partition. (Rank 2) |
| Viral Proteome | 14 | 18 | 77.7% | Quite high number of samples clustered. (Rank 3) |

### Principle Component Analysis:

This result indicates that the samples cluster in three distinct partitions.

TABLE III  
Cluster statistics from Principle Component Analysis

|  |  |  |
| --- | --- | --- |
| Sample | Principal Component | Interpretation |
| Eukaryotic Proteome | [0.48020417, 0.29809942] | Three distinct Principal Components imply three distinct classifications of the vectors. |
| Bacterial Proteome | [0.80249911, 0.08815054] |
| Viral Proteome | [0.43344298, 0.15875654] |

### Conclusion for Hypothesis 1:

The results obtained through the said analysis all agrees to our initial hypothesis that the Evolutionary taxonomy, which is based on phylogenetic relationship, can be revalidated through the proteomic sequences. This comes from the result that the proteomes of different classes of organisms and viruses clustered together, leading to a clear distinction between different phylogenetic branches, agreeing to the evolutionary taxonomy.

## (Hypothesis 2)

1. Plots

### k-Mean Clustering:

The partitions obtained from the k-Mean clustering resulted in following for the data:

TABLE IV  
Statistics from k-Mean Clustering (96 samples)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Sample | Clustered | Total | Ratio | Interpretations |
| Bacterial Proteome | 33 | 48 | 68.7% | All bacteria do not get clustered in the same partition. |
| Viral Proteome | 32 | 48 | 66.6% | No proper clustering of viruses. |

The above proves that the data is too similar to be divided into two proper classes and hence our hypothesis can be assumed to be true to an accuracy of approximately 32.3%. Hence, there is a probability that we can map a virus to its host by knowing only its proteome sequence up to an accuracy of 32.3%. Increasing the sample length of bacteria to 250 does establish the hypothesis and increases the accuracy to 42%. Further increasing it to 692, an accuracy of 43% can be achieved.

The result proves our hypothesis wrong. This means that one cannot correctly map a virus to its host based only upon the proteome sequence of the two.

TABLE V  
Statistics from k-Mean Clustering (50 samples)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Sample | Clustered | Total | Ratio | Interpretations |
| *Homo sapiens* | 1 | 1 | 100% | Expected result for a single point data. |
| Virus that attack Human | 30 | 49 | 61.2% | 28 out of 48 viruses get clustered and tagged into a partition. |

Similar result can be observed here with only 38.78% accuracy in predicting the host of a virus.

TABLE VI  
Statistics from k-Mean Clustering (96 Samples)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Sample | Clustered | Total | Ratio | Interpretations |
| Virus that attack Bacteria | 24 | 48 | 50% | No proper clustering. |
| Virus that attack Human | 28 | 48 | 58.3% | 28 out of 48 viruses get clustered and tagged into a partition. |

The above result shows that no clear difference can be established between the viruses that have different hosts. Hence, proteome sequence cannot be the only measure to predict and tag viruses for their hosts.

## (Hypothesis 3)

The average of the number of occurrences of the sequences in the proteome was found out to be 1.91 and its variance to be 12.85. This suggested that there could be some sequences that occurred multiple times. Following were the sequences with highest frequency of occurrences in descending order:

TABLE VII  
Statistics from k-Mean Clustering (96 Samples)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 1614 | 679 | 671 | 654 | 494 |
| GGAGG | GGNGG | AGGAG | GAGGA | GGTGG |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 381 | 293 | 284 | 282 | 275 |
| GGDGG | NGGAG | GNGGA | GNGGN | AAAA |

Trying to find the presence of these sequences in known motifs of Mycobacterium Tuberculosis could not produce any concrete result. Though, looking at the above table we can conclude that the sequences are Glycine (G) rich. Referring to the paper we can see that these Glycine rich proteins are in fact important as they regulate the cell outer structure (antigen) of this particular bacteria.

Moreover, by Table VII, relative frequency of Glycine in secondary structures infer that it is predominantly found in turns, suggesting, Glycine must be an important amino acid which modulates the structure of the proteins in *Mycobacterium Tuberculosis*.

1. Section Headings

No more than 3 levels of headings should be used. All headings must be in 10pt font. Every word in a heading must be capitalized except for short minor words as listed in Section III-B.

1. Level-1 Heading: A level-1 heading must be in Small Caps, centered and numbered using uppercase Roman numerals. For example, see heading “III. Page Style” of this document. The two level-1 headings which must not be numbered are “Acknowledgment” and “References”.
2. Level-2 Heading: A level-2 heading must be in Italic, left-justified and numbered using an uppercase alphabetic letter followed by a period. For example, see heading “C. Section Headings” above.
3. Level-3 Heading: A level-3 heading must be indented, in Italic and numbered with an Arabic numeral followed by a right parenthesis. The level-3 heading must end with a colon. The body of the level-3 section immediately follows the level-3 heading in the same paragraph. For example, this paragraph begins with a level-3 heading.
4. Figures and Tables

Figures and tables must be centered in the column. Large figures and tables may span across both columns. Any table or figure that takes up more than 1 column width must be positioned either at the top or at the bottom of the page.

Graphics may be full color. All colors will be retained on the CDROM. Graphics must not use stipple fill patterns because they may not be reproduced properly. Please use only *SOLID FILL* colors which contrast well both on screen and on a black-and-white hardcopy, as shown in Fig. 1.

gv_figure_4

Fig. 1 A sample line graph using colors which contrast well both on screen and on a black-and-white hardcopy

Fig. 2 shows an example of a low-resolution image which would not be acceptable, whereas Fig. 3 shows an example of an image with adequate resolution. Check that the resolution is adequate to reveal the important detail in the figure.

Please check all figures in your paper both on screen and on a black-and-white hardcopy. When you check your paper on a black-and-white hardcopy, please ensure that:

* the colors used in each figure contrast well,
* the image used in each figure is clear,
* all text labels in each figure are legible.

1. Figure Captions

Figures must be numbered using Arabic numerals. Figure captions must be in 8 pt Regular font. Captions of a single line (e.g. Fig. 2) must be centered whereas multi-line captions must be justified (e.g. Fig. 1). Captions with figure numbers must be placed after their associated figures, as shown in Fig. 1.



Fig. 2 Example of an unacceptable low-resolution image



Fig. 3 Example of an image with acceptable resolution

1. Table Captions

Tables must be numbered using uppercase Roman numerals. Table captions must be centred and in 8 pt Regular font with Small Caps. Every word in a table caption must be capitalized except for short minor words as listed in Section III-B. Captions with table numbers must be placed before their associated tables, as shown in Table 1.

1. Page Numbers, Headers and Footers

Page numbers, headers and footers must not be used.

1. Links and Bookmarks

All hypertext links and section bookmarks will be removed from papers during the processing of papers for publication. If you need to refer to an Internet email address or URL in your paper, you must type out the address or URL fully in Regular font.

1. References

The heading of the References section must not be numbered. All reference items must be in 8 pt font. Please use Regular and Italic styles to distinguish different fields as shown in the References section. Number the reference items consecutively in square brackets (e.g. [1]).

When referring to a reference item, please simply use the reference number, as in [2]. Do not use “Ref. [3]” or “Reference [3]” except at the beginning of a sentence, e.g. “Reference [3] shows …”. Multiple references are each numbered with separate brackets (e.g. [2], [3], [4]–[6]).

Examples of reference items of different categories shown in the References section include:

* example of a book in [1]
* example of a book in a series in [2]
* example of a journal article in [3]
* example of a conference paper in [4]
* example of a patent in [5]
* example of a website in [6]
* example of a web page in [7]
* example of a databook as a manual in [8]
* example of a datasheet in [9]
* example of a master’s thesis in [10]
* example of a technical report in [11]
* example of a standard in [12]

1. Conclusions

The version of this template is V2. Most of the formatting instructions in this document have been compiled by Causal Productions from the IEEE LaTeX style files. Causal Productions offers both A4 templates and US Letter templates for LaTeX and Microsoft Word. The LaTeX templates depend on the official IEEEtran.cls and IEEEtran.bst files, whereas the Microsoft Word templates are self-contained. Causal Productions has used its best efforts to ensure that the templates have the same appearance.

Causal Productions permits the distribution and revision of these templates on the condition that Causal Productions is credited in the revised template as follows: “original version of this template was provided by courtesy of Causal Productions (www.causalproductions.com)”.

1. Acknowledgment

The heading of the Acknowledgment section and the References section must not be numbered.

Causal Productions wishes to acknowledge Michael Shell and other contributors for developing and maintaining the IEEE LaTeX style files which have been used in the preparation of this template. To see the list of contributors, please refer to the top of file IEEETran.cls in the IEEE LaTeX distribution.

1. References
2. S. M. Metev and V. P. Veiko, *Laser Assisted Microtechnology*, 2nd ed., R. M. Osgood, Jr., Ed. Berlin, Germany: Springer-Verlag, 1998.
3. J. Breckling, Ed., *The Analysis of Directional Time Series: Applications to Wind Speed and Direction*, ser. Lecture Notes in Statistics. Berlin, Germany: Springer, 1989, vol. 61.
4. S. Zhang, C. Zhu, J. K. O. Sin, and P. K. T. Mok, “A novel ultrathin elevated channel low-temperature poly-Si TFT,” *IEEE Electron Device Lett.*, vol. 20, pp. 569–571, Nov. 1999.
5. M. Wegmuller, J. P. von der Weid, P. Oberson, and N. Gisin, “High resolution fiber distributed measurements with coherent OFDR,” in *Proc. ECOC’00*, 2000, paper 11.3.4, p. 109.
6. R. E. Sorace, V. S. Reinhardt, and S. A. Vaughn, “High-speed digital-to-RF converter,” U.S. Patent 5 668 842, Sept. 16, 1997.
7. (2002) The IEEE website. [Online]. Available: http://www.ieee.org/
8. M. Shell. (2002) IEEEtran homepage on CTAN. [Online]. Available: http://www.ctan.org/tex-archive/macros/latex/contrib/supported/IEEEtran/
9. FLEXChip Signal Processor (MC68175/D), Motorola, 1996.
10. “PDCA12-70 data sheet,” Opto Speed SA, Mezzovico, Switzerland.
11. A. Karnik, “Performance of TCP congestion control with rate feedback: TCP/ABR and rate adaptive TCP/IP,” M. Eng. thesis, Indian Institute of Science, Bangalore, India, Jan. 1999.
12. J. Padhye, V. Firoiu, and D. Towsley, “A stochastic model of TCP Reno congestion avoidance and control,” Univ. of Massachusetts, Amherst, MA, CMPSCI Tech. Rep. 99-02, 1999.
13. Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specification, IEEE Std. 802.11, 1997.

1. 17 viruses sampled randomly from 429 Viruses. [↑](#footnote-ref-1)
2. 18 bacteria, sampled randomly from 629 Bacteria. [↑](#footnote-ref-2)