HI 193.1 Activity: Recreation of *Generating Phenograms using Frequent Structure Mining Over Metabolic Pathways*

First Author#1, Second Author\*2, Third Author#3

*#First-Third Department, First-Third University  
Address Including Country Name*

1first.author@first-third.edu

3third.author@first-third.edu

*\*Second Company  
 Address Including Country Name*

2second.author@second.com

***Abstract*—The purpose of this activity is to recreate Osorio's paper on phenogram generation to explore the topological similarities of different taxa using their glycolysis and citrate cycle networks. The group presents the phenograms generated using the Jaccard Similarity Indices and the Hamming Distance. These are then compared with NCBI generated phenograms using Phylip branch matchings and Maximum Consensus Tree.**

***Keywords***— **phenogram,**

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. Research Methods

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. *Constructing FP-Table*

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. Constructing FP-Table

Using Python, FP-Table is generated from the created sets of taxa.

1. *Creating the table*

An empty list of distinct edges is initialized to serve as a container for the combined edges (tuples) of all taxa in a specific set of taxa. To preserve its uniqueness, duplicate edges are removed. This characterizes the frequent subgraph of a particular graph dataset G.

Let G = {G1, G2, ... Gn}, n = 5, be the set of taxa converted into digraphs. DE is a list of tuples of ordered pairs (A,B) representing distinct edges from the combined edges of Gk.

1. *Assigning bit codes*

Bit codes, say m, are assigned for each distinct edge DEi. Using the bitCheck() method, a value of 1 is returned to indicate the presence of DEi in a particular graph Gj, otherwise a value of 0 is returned. The list of bit codes has a k-length bit. In this case, the bit code list has a length of 5, representing the 5 chosen taxa.

Additionally, upon completing the assignment of bits for each representation of taxa by a distinct edge DE, a temporary element named “parity” with a value corresponding to the frequency of 1’s in a list, is appended at the end of each bit code list for easier sorting.

1. *Sorting FP-Table*

For convenient classification, distinct edges DE are ordered based on the bit code weight determined by the number of 1’s (parity) and its decimal values. Initially, distinct edges DE are classified based on the temporarily appended element parity. It is arranged in descending order of parities. This denotes the clustering of the edges. Secondly, distinct edges DE are sorted based on the decimal-value location of 1’s which are also decreasingly arranged. This represents the classification of the nodes.

Using Python, decimal values for each bit code list are temporarily appended at the end of the list for easy sorting. Both the temporary elements are removed immediately afterwards. Processes involved occur in the method sortTable().

1. *Generating the FP-Table*

The method generateFPTable() is called to return and display the table. It accepts a parameter of type list which contains the list of digraphs (class Taxa) possessing lists of vertices and edges. It is followed by passing the table to sortTable() for the classification and clustering of nodes and clusters. To sum up, the steps followed are specified below:

1.Get the set of taxa converted into digraphs.

2. Obtain distinct edges from all taxa in the set.

3. Derive the bit codes for each distinct edge.

4. Sort the edges according to the weight of the bit codes.

a. Sort by parity to classify Clusters.

b. Sort by decimal values of 1’s to classify Nodes.

1. Deriving the similarity index matrix

Using Python, FP-Table is generated from the created sets of taxa.

1. *Using Jaccard Similarity Index (JSI)g the table*

An empty list of distinct edges is initialized to serve as a container for the combined edges (tuples) of all taxa in a s

1. *Using Hamming Distance (HD)*

An empty list of distinct edges is initialized to serve as a container for the combined edges (tuples) of all taxa in a s

1. Results

After obtaining the phenograms from SIM using JSI and HD having inputted of two sets of taxa containing five species each as indicated in the following table.

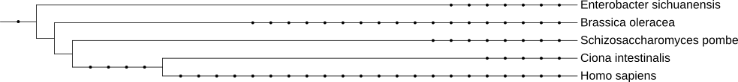
TABLE I  
selected species according to set

|  |  |
| --- | --- |
| **SET 1** | |
| **Code** | **Scientific name** |
| hsa | Homo sapiens |
| spo | Schizosaccharomyces pombe |
| cin | Ciona intesinalis |
| boe | Brassica oleracea |
| esh | Enterobacter sichuanensis |
| **SET 2** | |
| **Code** | **Scientific name** |
| thg | Vulcanisaeta distributa |
| hic | Haemophilus influenzae |
| ini | Ipomoea nil |
| ovi | Opisthorchis viverrini |
| psoj | Phytophthora sojae |

The following diagram is an NCBI-generated phenogram. This will be used as basis in comparing phenograms generated from SIM using JSI and HD in terms of matchings. Phylip branch matching and maximum consensus tree are the tools used to evaluate and further infer similarities of the compared phenograms.

FIGURE X  
ncbi-generated phenograms

Set 1



A picture containing diagram

Description automatically generatedSet 2

The % match using Phylip was computed using the following formula:

Text

Description automatically generated

Maximum consensus tree was also used in computing for the same attribute:

Text

Description automatically generated

TABLE 2  
summary of similarity comparisons

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **% match** | | | |
|  | **Set 1** | | **Set 2** | |
|  | **JSI** | **HD** | **JSI** | **HD** |
| **Phylip** |  |  |  |  |
| **MCT** | **1** | **1** | **8/7** | **8/7** |

hics 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.

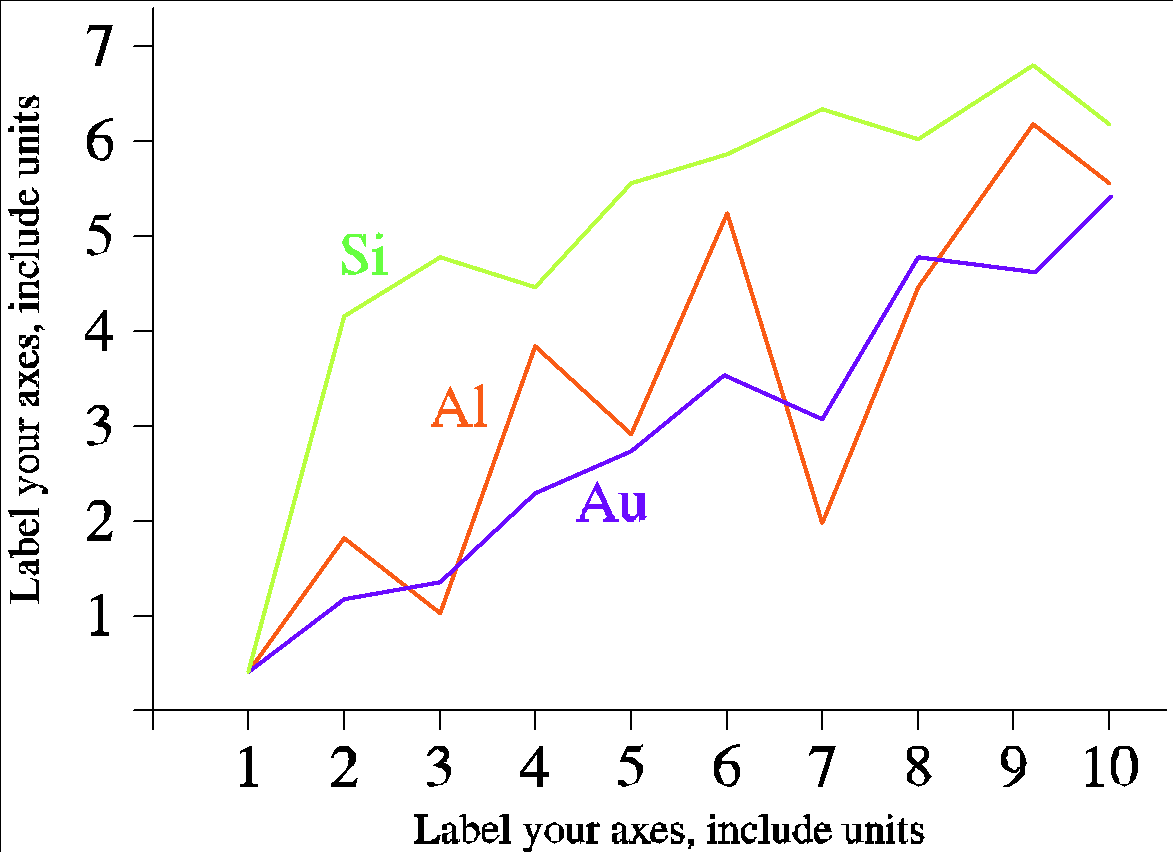


Fig. 1 A sample line graph using colors which contrast well both on screen and on a black-and-white hardcopy

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)”.

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.

References

1. S. M. Metev and V. P. Veiko, *Laser Assisted Microtechnology*, 2nd ed., R. M. Osgood, Jr., Ed. Berlin, Germany: Springer-Verlag, 1998.
2. 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.
3. 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.
4. 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.
5. 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.
6. (2002) The IEEE website. [Online]. Available: http://www.ieee.org/
7. M. Shell. (2002) IEEEtran homepage on CTAN. [Online]. Available: http://www.ctan.org/tex-archive/macros/latex/contrib/supported/IEEEtran/
8. *FLEXChip Signal Processor (MC68175/D)*, Motorola, 1996.
9. “PDCA12-70 data sheet,” Opto Speed SA, Mezzovico, Switzerland.
10. 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.
11. 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.
12. *Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specification*, IEEE Std. 802.11, 1997.