UNIVERSITI MALAYSIA SARAWAK

UNIMAS

Faculty of Cognitive Sciences and Human Development

Visual Perception Enhancement in Sequence Logo

Kok Weiying

Master of Science
2018

Visual Perception Enhancement in Sequence Logo

Kok Weiying

A thesis submitted

In fulfilment of the requirements for the degree of Master of Science

(Human Factors)

Faculty of Cognitive Sciences and Human Development
UNIVERSITI MALAYSIA SARAWAK
2018

UNIVERSITI MALAYSIA SARAWAK

Grade: ____________

Please tick (√)

Final Year Project Report □

Masters ☑

PhD □

DECLARATION OF ORIGINAL WORK

This declaration is made on the 27th day of September 2018

Student's Declaration:

I Kok Weiying, 13020101, Faculty of Cognitive Sciences & Human Development (PLEASE INDICATE STUDENT'S NAME, MATRIC NO. AND FACULTY) hereby declare that the work entitled, Visual Perception Enhancement in Sequence Logo is my original work. I have not copied from any other students' work or from any other sources except where due reference or acknowledgement is made explicitly in the text, nor has any part been written for me by another person.

27/9/2018

Date submitted

KOK WEIYING 13020101

Name of the student (Matric No.)

Supervisor's Declaration:

I OON YIN BEE (SUPERVISOR'S NAME) hereby certifies that the work entitled, VISUAL PERCEPTION ENHANCEMENT IN SEQUENCE LOGO (TITLE) was prepared by the above named student, and was submitted to the "FACULTY" as a * partial/full fulfillment for the conferment of MASTER OF SCIENCE (PLEASE INDICATE THE DEGREE), and the aforementioned work, to the best of my knowledge, is the said student's work

Received for examination by: Oon Yin Bee

(Name of the supervisor)

Date: 3/10/2018

I declare this Project/Thesis is classified as (Please tick (√)):

□ CONFIDENTIAL (Contains confidential information under the Official Secret Act 1972)*
□ RESTRICTED (Contains restricted information as specified by the organisation where research was done)*
☑ OPEN ACCESS

Validation of Project/Thesis

I therefore duly affirmed with free consent and willingness declared that this said Project/Thesis shall be placed officially in the Centre for Academic Information Services with the abide interest and rights as follows:

- This Project/Thesis is the sole legal property of Universiti Malaysia Sarawak (UNIMAS).
- The Centre for Academic Information Services has the lawful right to make copies for the purpose of academic and research only and not for other purpose.
- The Centre for Academic Information Services has the lawful right to digitise the content to for the Local Content Database.
- The Centre for Academic Information Services has the lawful right to make copies of the Project/Thesis for academic exchange between Higher Learning Institute.
- No dispute or any claim shall arise from the student itself neither third party on this Project/Thesis once it becomes sole property of UNIMAS.
- This Project/Thesis or any material, data and information related to it shall not be distributed, published or disclosed to any party by the student except with UNIMAS permission.

Student's signature ______________ (Date) 27/9/2018

Supervisor's signature: ______________ (Date) 3/10/2018

Current Address:
No. 2, Jalan 1/7, Taman Sri Kluang,
86000 Kluang, Johor.

Notes: * If the Project/Thesis is **CONFIDENTIAL** or **RESTRICTED**, please attach together as annexure a letter from the organisation with the period and reasons of confidentiality and restriction.

[The instrument was duly prepared by The Centre for Academic Information Services]

DECLARATION

I hereby declare that the work entitled “Visual Perception Enhancement in Sequence Logo” is my original work. I have not copied from any other student’s work or from any other sources except where due reference or acknowledgement is made explicitly in the text, nor has any part been written for me by another person. The thesis has not been accepted for any degree and is not concurrently submitted in the candidature of any other degree.

____

Name of Student: Kok Weiying

Date:

ACKNOWLEDGEMENT

I would like to dedicate my special thanks to my supervisor and co-supervisor Ms. Oon Yin Bee and Dr. Lee Nung Kion for their help and support with patience throughout this project. The encouragement, suggestions and guidance have really helped me a lot in completing this thesis. Besides, I would also like to thank my family members and friends for their love and support during my studies at UNIMAS. Without them, the journey of studying will be hard. I would also like to thanks all the participants who are willing to help me in completing the questionnaire and the students and lecturers from Faculty of Resource Science and Technology who are willing to share their knowledge and opinion with me throughout this research journey. Their good advice and suggestions have helped me a lot in completing this project. Lastly, I offer my regards and blessings to all of those who had help and supported me in any aspect throughout this project completion.

ABSTRACT

Sequence Logo is a popular graphical representation to visualize the conservation characteristics of biological sequence motifs profile. Previous studies have found that the used of sequence logo as a visual representation to support scientific evidence or arguments could cause misinterpretation by users with different levels of knowledge and experiences and consequently causes biases in decision making. This study is separated into two evaluation studies. The first study is the evaluation of the current design of the Sequence Logo. The aim of the study is to identify the differences in perception and decision making between users with diverse skills and experience level while visualizing Sequence Logo. Whereas the aim of the second study is to evaluate on the improved design of Sequence Logo to identify if the improvement by using different visual attributes will help diverse users in perceiving and interpreting the information to alleviate biases and misinterpretation of the results. In the first evaluation study, an online survey is carried out on a voluntary basis where 52 participants from bioinformatics, genetic or molecular biology background had involved in the survey. Paired sample t-test and independent t-test was used to analyze the results obtained. The result shows that there are significant differences in the perception and needs between novice and expert users while interpreting the results in Sequence Logo. Visual cues and detailed information display are needed by novice users whereas experts prefer a simple but more functional representation in the sequence logo. The second evaluation study involves 55 users with experience in using Sequence Logo. The result shows that the improvement on the colour will help the user in identifying the conservation level of the nucleotide, however, the arrangement and the amount of information present in the improved Sequence Logo causes more attention needed while perceiving the results. Therefore, enhancement on the improved design of the

Sequence Logo is needed in terms of the colour to represent the non-conserved nucleotide, the arrangement of the nucleotide and the amount of information shown on the graphical representation. Interactivity on the tools is also needed to help both novice and expert user in choosing the most suitable graphical representation for the analysis of result or the representation of result in publication.

Keywords: Sequence Logo, consensus sequence, Gestalt perception, novice, expert

Alat Visualisasi dan Analisis Motif Biologi Interaktif

ABSTRAK

'Sequence Logo' merupakan representasi grafik yang popular untuk memvisualisasikan ciri-ciri pemuliharaan profil motif urutan biologi. Kajian terdahulu telah mendapati bahawa penggunaan 'Sequence Logo' sebagai perwakilan visual sebagai bukti atau argumen saintifik boleh menyebabkan salah tafsir oleh pengguna dengan tahap pengetahuan dan pengalaman yang berbeza dan seterusnya menyebabkan bias dalam membuat keputusan. Kajian ini dibahagikan kepada dua kajian penilaian. Kajian penilaian pertama ialah penilaian reka bentuk 'Sequence Logo' yang sedia ada. Tujuan kajian ini adalah untuk mengenal pasti perbezaan persepsi dan pengetahuan antara pengguna dengan tahap kemahiran dan pengalaman yang berbeza semasa memvisualisasikan 'Sequence Logo'. Manakala tujuan kajian kedua adalah untuk menilai reka bentuk 'Sequence Logo' yang ditambah baik dengan mengikuti beberapa teori persepsi untuk mengenalpasti adakah peningkatan dengan menggunakan ciri visual yang berbeza akan membantu pengguna yang beragam dalam melihat dan menafsirkan maklumat untuk mengurangkan bias dan salah tafsir hasilnya. Dalam kajian penilaian pertama, kaji selidik dalam talian dijalankan secara sukarela di mana 52 peserta dari latar belakang bioinformatik, genetik atau molekul biologi terlibat dalam tinjauan. Ujian t-pasangan yang sepadan dan ujian t-bebas digunakan untuk menganalisis hasil yang diperolehi. Keputusan menunjukkan bahawa terdapat perbezaan yang signifikan dalam persepsi dan keperluan antara pemula dan pakar semasa mentafsir keputusan dalam 'Sequence Logo'. Petunjuk visual dan paparan maklumat terperinci diperlukan oleh pengguna pemula sedangkan pakar lebih suka perwakilan yang mudah tetapi lebih berfungsi dalam logo urutan. Kajian penilaian kedua melibatkan 55 pengguna yang

berpengalaman menggunakan ‘Sequence Logo’. Keputusan menunjukkan bahawa penambahbaikan pada warna akan membantu pengguna dalam mengenal pasti tahap pemuliharaan nukleotida namun susunan dan jumlah maklumat yang terdapat dalam ‘Sequence Logo’ yang lebih terlalu banyak akan menyebabkan lebih banyak perhatian diperlukan ketika melihat hasilnya. Oleh itu, peningkatan pada reka bentuk ‘Sequence Logo’ yang lebih baik diperlukan dari segi warna untuk mewakili nukleotida yang tidak konservasi, susunan nukleotida dan jumlah maklumat yang ditunjukkan pada perwakilan grafis. Interaktiviti pada alat juga diperlukan untuk membantu kedua-dua pemula dan pengguna pakar dalam memilih perwakilan grafis yang paling sesuai untuk analisis hasil atau perwakilan hasil penerbitan.

Kata kunci: *‘Sequence Logo’, urutan consensus, persepsi Gestalt, orang baru, pakar*

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION

1.1 Overview

This chapter discusses the background of research, problem statement, objectives, the scope of the project, the significance of the research and summary of the chapter. The background of the study describes the importance of sequence motif discovery and visualization in the biological field and how visualization can influence the user's interpretation of the information. The problem statement rationalizes the objectives and the significance of the study and the final section summarize this chapter.

1.2 Background of the Study

Software for discovering and visualizing sequence motifs are essential tools for life scientists in solving various motif discovery problems. The growths of motif databases such as JASPAR, SCPD, TRANSFAC, PROSITE, PRINTS, and etc. for DNA and protein motifs have made the motif discovery an important computational problem in discovering the sequence patterns (Bailey, 2008).

Biological sequence motifs are short recurring sequence patterns that represent many features in DNA, RNA, and proteins. It can represent the DNA transcription factor binding sites (TFBS), the RNA splice junction or the protein molecules binding domains (Lin, 2012). In DNA, TFBS motifs help to specify the order and nucleotide preference for a particular TF at each position of the binding site (Bailey, 2008). Protein motifs can represent the active site of enzymes and categorise the protein regions that are involved in determining the protein structure and stability (Bailey, 2008). Challenges were faced by researchers in identifying the motifs as the sequence motifs are never exactly the same as the actual conserved sequence and the regulatory sequences that contain the motifs are sometimes located very far away from the coding regions (Chauhan & Agarwal, 2012). Therefore, discovering and visualizing sequence motifs are very important in the biological field as it will lead to a better understanding of transcription regulation, splicing of mRNA, and the formation of protein complexes (Bailey, 2008).

Data visualization tools are important in providing valuable complements to the automated computational techniques which enable researchers to derive scientific insight from the large-scale of biological data. It can augment our ability to give reasons for a complex data and furthermore helps to increase the efficiency of the analysis of data. Sequence Logo is one of the popular visualization methods for displaying the conservation characteristics of a sequence motif profile obtained from wet-lab or through computational analysis (Schneider & Stephens, 1990). A Sequence Logo illustrates the information about the motif conservation characteristics and the relative frequency of the nucleotide or amino acid with different character symbols and size. Figure 1.1 shows an example of the Sequence Logo of the E. coli transcription factor binding sites.

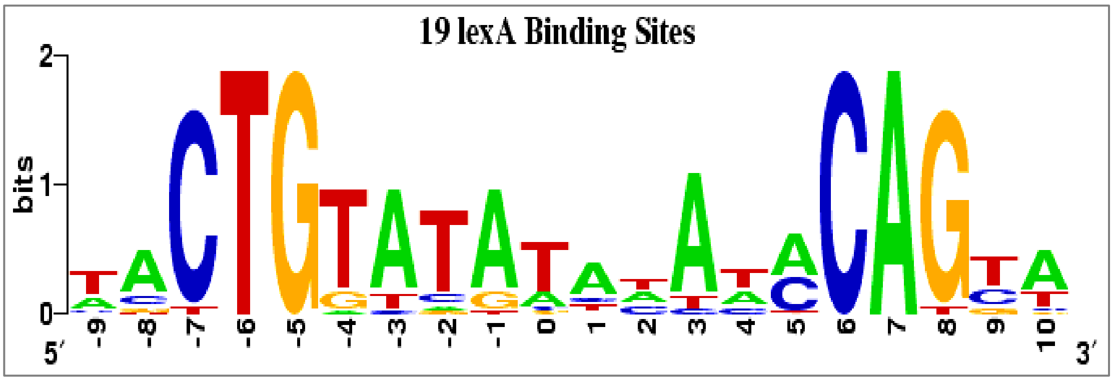

Figure 1.1: An example of a motif Sequence Logo of E. coli transcription factor binding sites.

The primary goal for the graphical representation is to clearly and efficiently transform the huge amount of data into simple lines and bars to help the user to easily analyse and give reasons or meaning for the representation (Few, 2004). Gestalt Principles of perception is one of the earliest contributions to the study of perception to uncover the way the visual system perceive pattern, form, and organize the things we see. It was found that the visual system will organize what we see in particular ways in an effort to make sense of it (Few, 2013). Knowledge about human perception is greatly important in designing an effective graphical representation in order to understand how human tends to perceive the attributes inside a graphical representation. The human visual system can easily distinguish the difference in size, colour, shape, length, and orientation. This is called pre-attentive processing where the detection of the difference for these pre-attentive attributes can be done without any significant processing effort (Few, 2004). An effective graphical representation will take advantage of the pre-attentive attributes to show comparison or convey messages to the user for decision making (Few, 2004).

Data visualization is effective because it helps to shifts the balance between perception and cognition to take fuller advantage of the brain's abilities in perceiving and making sense of the things perceived. Therefore, the cognition of the user plays an important role in how the graphical representation is being perceived. Cognition refers to the mental process involving perception, attention, memory which will assist us to remember, think, solve the problem, and make a decision (Difference Between Cognition and Perception, 2014). Individuals with different level of knowledge, experience, and skills will have different perception and their decision-making ability are build up from these experience and practices (Randel, Pugh, & Reed, 1996). Previous studies have found that there is a difference between the user with less experience and knowledge about the domain (novice user) and user with more experience and knowledge on the domain (expert users) explore data visualizations (Zhu, 2007). Novice users are unable to utilize visual cues in a graphical representation where novice users tend to confuse the visibility with relevance whereas expert user is able to disregard the irrelevant information and match patterns based on their knowledge and experience (Petre & Green, 1993).

1.3 Problem Statement

The exponentially increasing amount of biological data is challenging the abilities of biological scientists in making sense of all the data in a concise and meaningful way (O'Donoghue, et al., 2010). Nowadays, computer-based visualization has been widely used by biologist to understand and communicate data, to generate ideas and gain insight into biological processes. Compared to twenty years ago where only experts are able to create visual representation of a protein structure, or a large phylogenetic tree, the advancement of computer hardware and network has increased the accessibility and the use of visualization software where many visualization tasks can now be easily managed by user with a standard personal computer (O'Donoghue, et al., 2010). However, the diversity and a large number of tools available can make the problem of visualizing worse and confusing. Several issues that are widely addressed by biologists on these visualization tools are:

a) Standardize graphical representation

- The lack of standards in representation is one of the problems faced by end users due to the rapid evolving of the visualization method. Although diversity and innovated graphical representation are needed in the design of the visualization, the improvement on the ease of use of the tools can be largely enhanced by adopting some standards in representation (O'Donoghue, et al., 2010).

b) Visual Analytics

- Finding a balance between visualization and functionality of a tool is a challenge. Visual analytics methods which involve the studying of visualization in the process of analysing and understanding data is important to improve the ability of tools to provide meaningful biological insights (O’Donoghue, et al., 2010).

c) Ease of use

- Biologists usually fail to fully benefit from visualization methods because software tools are too difficult to learn especially for the novice user. Although there have been many advancements on understanding the underlying principles of usability being adopted by developers these improvements are still very slow, as the work on usability is usually less rewarded in the biological field than is in research on new methods (O’Donoghue, et al., 2010).

These issues show that improvement is needed by adopting some standards or framework when creating a biological visualization to reason about the spectrum and considerations to help scientist match their visualization goals with appropriate design consideration.

The graphical representation that will be focused on this study is Sequence Logo. It is a graphical representation that is widely used in many scientific studies related to the transcription analysis such as: (i) to evaluate the newly proposed algorithm or tools for solving sequence motif discovery problems; (ii) as the support evidence for the computational framework or the biological methodology; and (iii) used to compare the characteristics of the different binding site specification of a same transcription factor (TF)

(Lee & Oon, 2012). Nevertheless, several literatures have found some limitations in the Sequence Logo visualization that will affect the interpretability of the user. The well-known issues found on the Sequence Logo are:

(a) Issues on the design of the graphical representation

   i. The design of the representation such as the arrangement of the nucleotide, the arrangement of the consensus sequence, the colour, size, or shape of the nucleotide on the Sequence Logo is found to be misleading to the user. Previous studies have found that misinterpretation will occur because the arrangement for each conserved nucleotide in the Sequence Logo are assumed to be mutually independent with the same background distribution at each position as the design does not provide any indication of correlation between different position of the alignment (Bindewald, Schneider, & Shapiro, 2006; Vacic, Lakoucheva, & Radivojac, 2006).

   ii. The shape and size representation of the nucleotide A, C, G, and T will cause the assumption where every symbol was assumed to be equally distributed at each position (Li, et al., 2008).

   iii. The design of the graphical representation focus more on the conserved nucleotide where not much information about the under-represented symbols are shown on the logo (Li, et al., 2008). The stacking of the symbols can lead to confusion as the representation do not display any information about the nucleotide that is not present from the alignment columns and also lack of representation for gap symbols (Roca A. , 2014).