



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

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
DECLARATION.....	i
ACKNOWLEDGEMENT.....	ii
ABSTRACT.....	iii
ABSTRAK.....	v
TABLE OF CONTENTS.....	vii
LIST OF TABLES.....	xi
LIST OF FIGURES.....	xiii
CHAPTER 1: INTRODUCTION.....	1
1.1 Overview.....	1
1.2 Background of the Study.....	1
1.3 Problem Statement.....	4
1.4 Objectives.....	7
1.4.1 General Objectives.....	7
1.4.2 Specific Objectives.....	8
1.5 Research Framework.....	8
1.6 Research Hypotheses.....	9
1.6.1 The visual attributes performance in representing the information on Sequence Logo.....	9
1.6.2 The difference between novice and expert while interpreting Sequence Logo	10
1.6.3 The user’s perception of the design of the original and improved graphical representation.....	11

1.6.4 The perception of the improved Sequence Logo design based on the user's skill.....	12
1.7 Significant of the Study.....	13
1.8 Scope of the Study.....	15
1.9 Summary.....	16
CHAPTER 2: LITERATURE REVIEW.....	17
2.1 Overview.....	17
2.2 Sequence Logo.....	17
2.3 Previous Studies on Sequence Logo.....	19
2.4 Visualization.....	25
2.5 Perception.....	29
2.5.1 Gestalt Principle of Perception.....	30
2.6 Attention.....	38
2.7 Novice vs Expert.....	41
2.8 Summary.....	43
CHAPTER 3: EVALUATION STUDY ON CURRENT SEQUENCE LOGO	45
3.1 Overview.....	45
3.2 Research Method.....	45
3.2.1 Research Design.....	45
3.2.2 Population.....	47
3.2.3 Research Instrument.....	47
3.2.4 Procedure for Data Collection.....	49
3.2.5 Descriptive Statistics.....	49
3.2.6 Inferential Statistics.....	49

3.3 Results.....	50
3.3.1 Demographic Information.....	49
3.3.2 Close Ended Question Evaluation Results.....	52
3.3.3 Open Ended Question Results.....	73
3.4 Discussion.....	75
3.5 Summary.....	83
CHAPTER 4: EVALUATION STUDY ON THE IMPORVED SEQUENCE	
LOGO.....	84
4.1 Overview.....	84
4.2 Improvement Design on the Sequence Logo.....	84
4.3 Research Method.....	90
4.3.1 Research Design.....	90
4.3.2 Population.....	91
4.3.3 Research Instrument.....	92
4.3.4 Procedure for Data Collection.....	93
4.3.5 Descriptive Statistics.....	93
4.3.6 Inferential Statitics.....	93
4.4 Results.....	94
4.4.1 Demographic Information.....	94
4.4.2 Evaluation Based On User’s Perception Of The Design Of The Original and Improved Graphical Representation.....	96
4.4.3 Evaluation on The Perception of the Improved Sequence Logo Design Based On User’s Skill	108
4.5 Discussion.....	117

4.6 Summary.....	120
CHAPTER 5: CONCLUSION AND FUTURE WORKS.....	121
5.1 Overview.....	121
5.2 Summary of the Findings.....	121
5.3 Limitation of the Study.....	124
5.4 Future Works.....	125
5.5 Contribution of Data Visualization.....	126
REFERENCES.....	128
APPENDICES.....	133

LIST OF TABLES

	Page
Table 2.1 List of specialized or improved graphical representation similar with Sequence Logo.....	20
Table 2.2 Limitations found on the Sequence Logo.....	22
Table 2.3 Dreyfus novice expert transition characteristics.....	42
Table 2.4 Novice to Expert perception of context.....	43
Table 3.1 7 score Likert scale.....	48
Table 3.2 Demographic results.....	50
Table 3.3 Paired sample t-test for the arrangement and colour of the conserved nucleotide.....	55
Table 3.4 Paired sample t-test for the arrangement of the stack of nucleotide at each position in Sequence Logo.....	57
Table 3.5 Paired sample t- test on the ease to determine the relative conservation value of each nucleotide in the Sequence Logo.....	59
Table 3.6 Paired on the performance between Tool 1 & Tool 2 with the annotated motif.....	60
Table 3.7 Level of skills based on Dreyfus novice expert scale.....	62
Table 3.8 Frequency & Percentage of analytic & holistic group user.....	63
Table 3.9 Independent t-test result on the assess of context.....	64
Table 3.10 Frequency & percentage of rational & intuitive group user.....	69
Table 3.11 Independent t-test for the representation of information to assist in decision making between two groups.....	69
Table 4.1 7 score Likert scale.....	93

Table 4.2	Demographic results.....	94
Table 4.3	Paired sample t-test on the function of colour in the original and improved Sequence Logo.....	97
Table 4.4	The preference on the use of colour in the original and improved.....	99
Table 4.5	Paired sample t-test on the arrangement of the nucleotide in the original and improved Sequence Logo.....	100
Table 4.6	Paired sample t-test on the amount of information in the original and improved Sequence Logo.....	104
Table 4.7	Paired sample t-test on the user's preference of graphical representation.	106
Table 4.8	Level of skills based on Dreyfus novice expert scale.....	108
Table 4.9	Frequency & Percentage of analytic & holistic group user.....	109
Table 4.10	Independent t-test of how the colour of the nucleotide in the improved Sequence Logo is being assessed.....	109
Table 4.11	Independent t-test of how the arrangement of the nucleotide in the improved Sequence Logo is being assessed.....	111
Table 4.12	Independent t-test of how the information is being assess in the improved Sequence Logo.....	113
Table 4.13	Independent t-test of how analytic and holistic group user will use the improved Sequence Logo to represent the result.....	117

LIST OF FIGURES

	Page
Figure 1.1 An example of a motif Sequence Logo of E. coli transcription factor binding sites.....	2
Figure 1.2 Research Framework.....	9
Figure 2.1 CAP binding sites Sequence Logo.....	19
Figure 2.2 The DNA-binding helix-turn-helix motif of the CAP family.....	19
Figure 2.3 Comparison of the PWM-based algorithm evaluation results using Sequence Logo and Frequency Logo.....	24
Figure 2.4 Example application of law of Pragnanz in the arrangement of bar graphs	31
Figure 2.5 Example application of law of similarity in the design of bar graphs.....	32
Figure 2.6 Example application of law of proximity in the design of bar graphs.....	33
Figure 2.7 Example application of the law of continuity in the arrangement of bar Graphs.....	34
Figure 2.8 Example application of the law of common fate in the line graph.....	35
Figure 2.9 Example application of law of figure and ground in the design of bar graph.....	36
Figure 2.10 Example application of law of focal point in the design of bar graph.....	37
Figure 2.11 Example application of law of isomorphic correspondence in the design of bar graph.....	38
Figure 3.1 Procedure for data collection.....	49
Figure 3.2 Tools to generate Sequence Logo.....	52
Figure 3.3 Original arrangement and colour of the conserved nucleotide in Sequence Logo.....	53

Figure 3.4	Recommended arrangement and colour of the conserved nucleotide of the Sequence Logo – Same colour of the conserved nucleotide.....	54
Figure 3.5	Recommended arrangement and colour of the conserved nucleotide of the Sequence Logo – Bold alphabets for the conserved nucleotide.....	54
Figure 3.6	Empty spaces in the immediate adjacent position of every stack of nucleotide.....	56
Figure 3.7	Thin line separating every stack of nucleotide adjacent to each other.....	56
Figure 3.8	Vertical lines on the background indicating the vertical axis bit values.....	58
Figure 3.9	Bit values shown when the cursor is hovered over the nucleotide.....	58
Figure 3.10	(a) Annotated Motif; (b) Motif sequence logo predicted by Tool 1; (c) Motif sequence logo predicted by Tool 2.....	60
Figure 4.1	Sequence Diversity Diagram.....	84
Figure 4.2	Sequence Bundles.....	86
Figure 4.3	Sequence Logo with Glyph based.....	87
Figure 4.4	NHLH1 Sequence Logo.....	88
Figure 4.5	Improved Sequence Logo design.....	88
Figure 4.6	Improved Sequence Logo design with bits value.....	89
Figure 4.7	Procedure for data collection for the improved Sequence Logo evaluation	93
Figure 4.8	Tools to generate Sequence Logo.....	96

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.

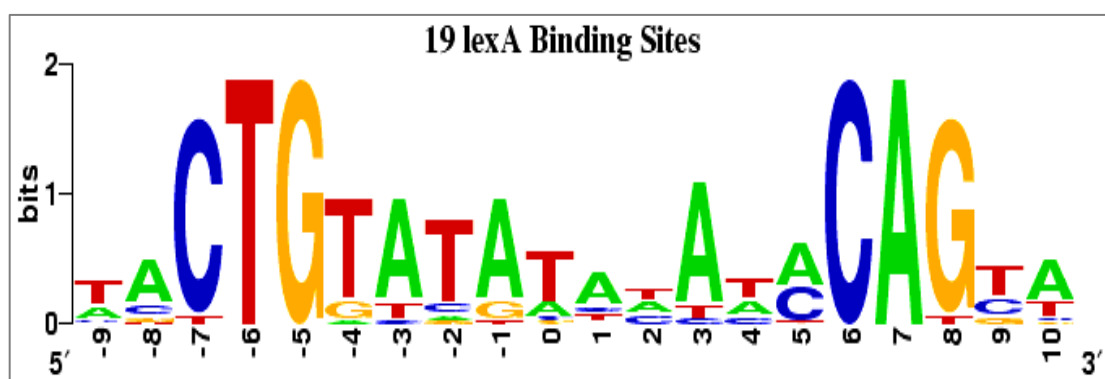


Figure 1.1: An example of a motif Sequence Logo of *E. coli* transcription factor binding sites.

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).

- iv. The colour of the nucleotide that are not standardized will cause confusion to the user when different colour is used to represent the nucleotide for different Sequence Logo (Lee & Oon, 2013).

(b) Lack of information on the graphical representation

- i. The lack of information such as the number of sequence alignment used to generate the Sequence Logo causes user unable to compare between two Sequence Logo (Vacic, Lakoucheva, & Radivojac, 2006).
- ii. Interpretation problem occurs for users with different level of experience due to the limited amount of information on the Sequence Logo (Li, et al., 2008). Constraint on the perception and decision-making performances for user with different level of experience occur due to the difference amount of information needed to assist in decision making (Lee & Oon, 2012). Moreover, confusion to the user who has less experience in using Sequence Logo will occur while interpreting the results presented in publication especially the logo size are very small (Weirauch, et al., 2013).

1.4 Objectives

The objectives of the study are divided into two parts which are the general objective and the specific objectives.

1.4.1 General Objectives

The general objective of this research is to identify the problems found on Sequence Logo in terms of the design of the graphical representation such as arrangement, colour, size, shape, and the amount of information displayed on the graphical representation. Besides that, an improved Sequence Logo visualization method is proposed to standardize the graphical

representation and provide more information on the logo, to reduce misinterpretation of the results while visualizing Sequence Logo.

1.4.2 Specific Objectives

The specific objectives of this research are:

- i. To evaluate the visual attributes performance on the design of Sequence Logo based on the human factors principles in perception and visualization;
- ii. To identify the difference between novice and experts while interpreting the result from Sequence Logo based on the Dreyfus novice expert skill acquisition model;
- iii. Design an improved Sequence Logo graphical representation based on the principles of perception and visualization.
- iv. To evaluate the user's perception of the design of the original and improved graphical representation;
- v. To identify the difference in perception between novice and experts while interpreting the improved Sequence Logo design;

1.5 Research Framework

The research framework of this study is illustrated in Figure 1.2. The first phase of the study aimed to identify the issues found on the Sequence Logo in terms of the design of the graphical representation and the information that can be acquired from the representation by the user with different level of experience. The second phase of the study will evaluate on the proposed Sequence Logo design where improvement is made based on the problems found in the first phase of the study.

The independent variable in this study is the information and visual attributes on the graphical representation such as the arrangement, colour, size and shape of the logo whereas

the dependent variable in this study is the perception and skill level of the user which are a novice or expert user.

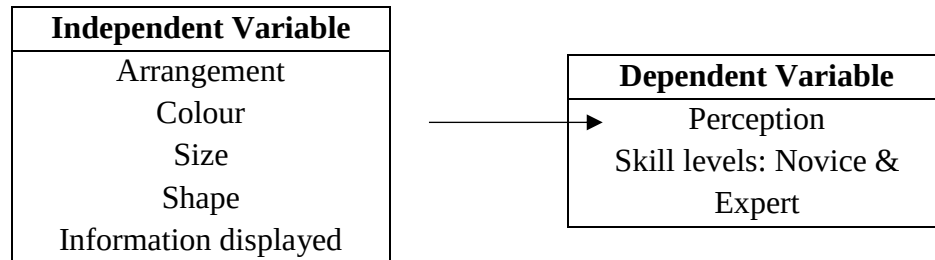


Figure 1.2: Research framework for the study

1.6 Research Hypotheses

This section generates the corresponding research hypothesis to the research objectives listed in Section 1.3.

1.6.1 The visual attributes performance in representing the information on Sequence Logo (Objective 1)

The corresponding research hypothesis for this objective is:

Hypothesis 1: There are some human factors issues found on the visual attributes representation of the Sequence Logo such as the colour, arrangement, shape and size which will affect the perception and interpretation of results on the graphical representation.

This research hypothesis intended to find out the issues on the visual representation of Sequence Logo which was said to be confusing and misinterpreting users. In order to find out the perception of the user on the design of the logo, a few improvements on the current design of Sequence Logo were created based on the Gestalt Principle of Similarity and Proximity. These improvements were used to compare with the original design to identify if these improvements made will help the user to better perceive the information on the logo compares to the original design. The null hypotheses generated were:

- H₀ 1:** There is no significant difference in the ease to identify the conserved nucleotide between the original arrangement and colour of the conserved nucleotide with the recommended arrangement and colour of the conserved nucleotide.
- H₀ 2:** There is no significant difference on the arrangement of the nucleotide by using the empty space or a thin line at the immediate adjacent position of every stack of the nucleotide to help in the grouping of the nucleotide at each position.
- H₀ 3:** There is no significant difference between ease to determine the relative conservation value of each nucleotide in the Sequence Logo with the vertical lines on the background and with the bit values shown when the cursor hovers over the nucleotide.
- H₀ 4:** There is no significant difference on how user perceives and rate the Sequence Logo result between Tool 1 & Tool 2 in terms of accuracy, the importance of nucleotide T at position 3, and the quality of the result.

These null hypotheses will help to identify the improvement that can be made on the logo in terms of colour, arrangement, size, shape, lines and amount of information to help users to perceive and interpret the results better. These improvements will then be used as a reference in designing the improved version of Sequence Logo.

1.6.2 The difference between novice and expert while interpreting Sequence Logo (Objective 2)

The corresponding hypothesis for this objective is:

Hypothesis 2: The different level of skills and experience will directly affect how the graphical representation are perceived and interpreted between novice and expert user.

This hypothesis intended to find out the difference between novice and expert user in perceiving the context in Sequence Logo and how the display of information will affect their decision-making ability. The null hypotheses generated were:

H₀ 1: There is no significant difference on how the context is being assessed in Sequence Logo between analytically (skills level 1-6) and holistically (skills level 7-10) group.

H₀ 2: There is no significant difference in the way rational thinking group (skills level 1-8) and intuitive thinking group (skills level 9-10) perceive the information in Sequence Logo to assist them in decision making.

These null hypotheses will help to identify the difference between novice and expert user in perceiving the context shown in Sequence Logo and how the perception of the information will affect their decision-making ability while interpreting the results. The problems identified will then be used to improve the design of the Sequence Logo for objective 3.

1.6.3 The user's perception of the design of the original and improved graphical representation (Objective 4)

The corresponding hypothesis for this objective is:

Hypothesis 3: The improved Sequence Logo design will help the user to better perceive and interpret the information compare to the original design.

This hypothesis intended to identify which improvement made based on the issues found from Objective 1 that will help the user to perceive and interpret the results better than the original design. The null hypotheses made were:

H₀ 1: There is no significant difference in how user perceives the function of colour between the original design and improved design of Sequence Logo.

H₀ 2: There is no significant difference in the preference of the user in using the colour to represent the information in the original and improved Sequence Logo.

H₀ 3: There is no significant difference on how the arrangement of the nucleotide in the original design and improved design of Sequence Logo helps the user to identify the consensus sequence in the representation.

H₀ 4: There is no significant difference on the information present in the original and improved Sequence Logo to help the user in analysing the results.

H₀ 5: There is no significant difference in the user's preference for using the original or improved Sequence Logo to represent the results.

These null hypotheses will help to identify which improvement of the visual attributes in terms of colour, arrangement, size and shape will help the user to perceive the information better compare to the original design. Besides that, the null hypothesis will also help to identify the user's preference for using the original or improved Sequence Logo to represent the result in their study.

1.6.4 The perception of the improved Sequence Logo design based on the user's skill (Objective 5)

The corresponding hypothesis for this objective is:

Hypothesis 4: The different level of skills and experience will directly affect how the graphical representation are perceived and interpreted by novice and expert user on the improved Sequence Logo design.

This hypothesis intended to identify the difference between novice and expert user in perceiving the context on the improved graphical representation and find out if these improvements will improve the decision-making process for all skill level user. The null hypotheses generated were:

H₀ 1: There is no significant difference on how the context of colour is being assessed in the improved Sequence Logo between analytically (skills level 1-6) and holistically (skills level 7-10) group.

H₀ 2: There is no significant difference on how the arrangement of the nucleotide is being assessed in the improved Sequence Logo between analytically (skills level 1-6) and holistically (skills level 7-10) group

H₀ 3: There is no significant difference on how the information display in the graphical representation is being assessed between analytically (skills level 1-6) and holistically (skills level 7-10) group.

H₀ 4: There is no significant difference on the way analytically (skills level 1-6) and holistically (skills level 7-10) group use the improved Sequence Logo as an extra reference to interpret the results but remain the use of Original Sequence Logo as the main graphical representation of the result.

These null hypotheses will help to identify which improvement on the graphical representation will help both novice and expert user to perceive and interpret the information better on the improved version of Sequence Logo.

1.7 Significant of the study

Visualization is not only explain how scientist makes sense of the data but also how they communicate with their findings. Techniques and guideline to create an effective

visualization are quite different depending on the goal to explain or explore however the spectrum of consideration is always being neglected while creating visualization (Görg, Hunter, Kennedy, O'Donoghue, & Van Wijk, 2012).

The tremendous growth of the bioinformatics field shows a great contribution to the field of medicine, biological research and computer science. Various tools were developed to help scientist and researchers to display and analyse a large amount of biological data in graphical representation however many of the tools developed are at their disposal as most of these tools are unable to be used by non-expert user (O'Donoghue, et al., 2010). User's previous knowledge is important as their intentions will influence their analytical reasoning and visual thinking (Sakai, 2016). Designing an effective visualization for a specialized domain is challenging as it requires the designer to acquire sufficient knowledge in the domain in order to understand the tasks of the user (Sakai, 2016). Although many tools were developed to assist biological scientist to visualize specific kinds of biological data, biologists still find that the information needed is not met by current tools and misinterpretation will occur during the analysis process and affect the accuracy of the reported results (O'Donoghue, et al., 2010). These problems were caused by the lack of a framework to reason about the spectrum and considerations to help scientists better match their visualization goals with appropriate design considerations (Görg et al., 2012).

Two fundamental problems with respect to the evaluation of tools for visualizing biological data are: (1) How to motivate biologists to participate in evaluations? and (2) How to evaluate the tools? In order to motivate biologist to participate in evaluation some benefits from the evaluation need to be given such as they might find a tool they can use to solve their problems (Görg et al., 2012). Whereas for the second question the gap between the designer working on the biological data visualization and designer working on information

visualization need to be close and combine as both approaches taken have their own merit in designing a visualization tool (Görg et al., 2012).

Therefore, it is important to understand both the needs of the biologist and the principles of visualization in order to close the gap between the groups and create a visualization that will benefit for all. Hence, this study will help to bridge the gap between both groups where the needs of the user (biologist) while using Sequence Logo is identified and the disciplines related to visualization and human perception are implemented into the design of the improved representation. Evaluation of the graphical representation based on the principles of visualization is carried out to provide insight on how the design of the visualization will affect the interpretation of information and decision-making process. Additionally, this study also highlights the important principles and theories in perception and visualization that can be used as a guideline for designing biological visualization tool so that the tools created are able to meet the needs of a biologist to convey the information effectively to the viewer.

1.8 Scope of the study

This study aims to find out the problems found in the design of Sequence Logo that causes misinterpretation to occur and provide an improved design that can help to solve the problem. The principles of perception and visualization such as the Gestalt principle of perception, attention and visualization were used to explain the way the visual attributes represent the information needed in the graphical representation. The function of each visual attribute such as colour, arrangement, shape, size, lines and white space were identified and evaluated to find out if these attributes able to convey correct information to the user in the graphical representation. Besides that, Dreyfus model of transition from novice to expert was used to find out the different needs of visual cue and information for a user with a

different level of experience in using the logo. The improved version of Sequence Logo was designed based on all the issues found from the first evaluation study with the reference of a few successful Sequence Logo design that was presented in the BioVis competition and some Sequence Logo designed by different biologist to solve different functionality problem according to their needs.

1.9 Summary

This chapter discussed the problems found in Sequence Logo based on previous studies, the objective of the research, the significance of the study and the scope involved in this study. In this study, the evaluation of the Sequence Logo is separated into two evaluation study where the first study is discussed in Chapter 3 which involved the evaluation of the current design of Sequence Logo to find out the problems found on the graphical representation and the discussion on the results found. The second study is discussed in Chapter 4 which involved an evaluation study of the improved Sequence Logo design based on the results from the first evaluation study. In the next chapter, previous studies on the Sequence Logo and related theories on human perception and visualization in the design of a graphical representation will be discussed.

CHAPTER 2

LITERATURE REVIEW

2.1 Overview

This chapter reviewed about the way Sequence Logo represents the sequence motifs, previous studies on the Sequence Logo graphical representation, the principles and theories on perception and visualization and the difference between novice and expert while perceiving a context.

2.2 Sequence Logo

Schneider and Stephens (1990) presented a graphical method for displaying the consensus sequences called Sequence Logos to help biologist in visualizing the conservation characteristic of a sequence motif for DNA, RNA or protein sequences. In Sequence Logos, individual nucleotides or characters representing the sequence are stacks on top of each other at each position where the height of each character is proportional to its frequency and the height of the stack is determined by the total information content. The characters at each position are sorted by positioning the most frequent character on top (Schneider & Stephens, 1990). A Sequence Logo concentrates on the information about the general consensus of the sequences, the order of predominance of the residue, the relative frequency of each residue, the amount of information present at each position and the initiation point or cut point of the sequences into a single graphical representation (Schneider & Stephens, 1990). It aims to aid user in the interpretation of the sequence data by visualizing the conserved motif that represents various functional or structural properties such as the transcription factor, the enzyme of DNA sequences, the proteolytic cleavage sites, the T-cell epitopes and the analysis of targets of neutralizing antibodies in HIV (Olsen et al., 2013).

In Sequence Logo, the information content present at each position of the multiple sequence alignments is calculated based on Shannon's information theory. Based on the information theory, the uncertainty measures are defined as:

$$H(l) = - \sum_{b=a}^t f(b,l) \log_2 f(b,l)$$

Where $H(l)$ is the uncertainty at position l , b is one of the bases in nucleotide or amino acid depend on the Sequence Logo that need to be generated, such as A, C, G or T in DNA sequence and $f(b,l)$ is the frequency of base b at position l . The total information content at a specific position is represented by the decrease in uncertainty as the binding site is aligned:

$$R_{sequence}(l) = \log_2 B - (H(l) + e(n))$$

Where $R_{sequence}(l)$ is the amount of information present at position l , $\log_2 B$ is the maximum uncertainty at any given position where B is the equiprobable characters ($\log_2 4 = 2$ bits for DNA/RNA and $\log_2 20 \approx 4.32$ bits for proteins) and $e(n)$ is a correction factor required when one only has a few sample sequences. The height of each character in the logo is calculated by multiplying the frequency of that base by the total information at that position:

$$Height(b,l) = f(b,l)R_{sequence}(l)$$

Based on the algorithms above, the bases are then stacked on top of each other with increasing frequency where the most abundant base at the topmost position.

After the Sequence Logos were introduced by Schneider and Stephens, several Sequence Logo generator is created by researchers and one of the most popular Sequence Logo generator is WebLogo (Olsen et al., 2013). WebLogo is created by Crooks et al. in

2004 that helps the user to generate standard Sequence Logos for DNA, RNA and protein sequences. Other than WebLogo, several specialized logo generators have been developed to assist the visualization of specific motifs or functional sequence units that are unapparent from the standard Sequence Logo (Olsen et al., 2013). Figure 2.1 and 2.2 shows an example of the standard graphical visualization of Sequence Logo for Catabolite Activator Protein created by WebLogo:

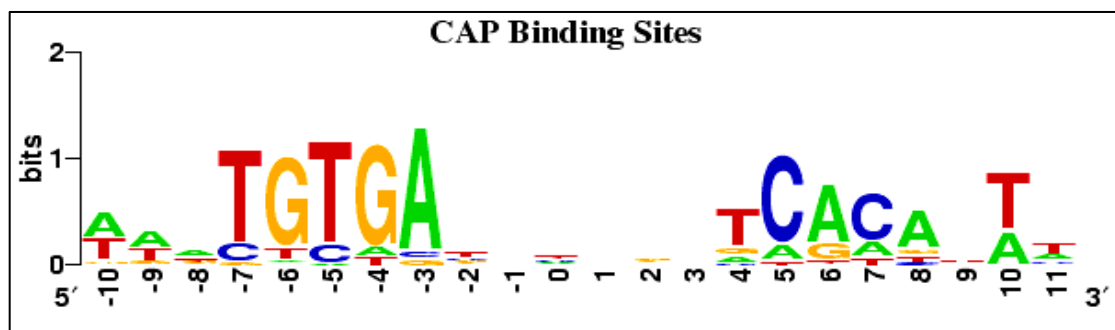


Figure 2.1: CAP binding sites Sequence Logo

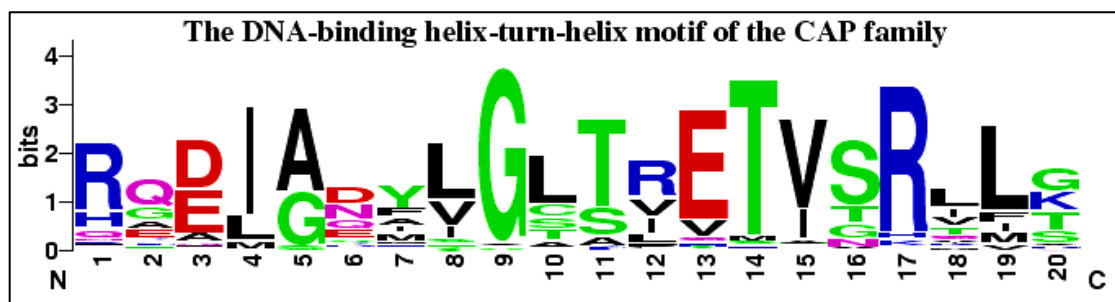


Figure 2.2: The DNA-binding helix-turn-helix motif of the CAP family

2.3 Previous studies on Sequence Logo

Currently, Weblogo (Crooks, Hon, Chandonia, & Brenner, 2004) is the most popular Sequence Logo generator which helps the user to generate standard Sequence Logo for DNA, RNA and protein sequences (Olsen et al., 2013). Besides that, several specialized or improved version of graphical representations have been developed to assist the visualization of specific motifs or functional sequence units that are not apparent in the standard Sequence Logo such as CorreLogo, enoLOGOS, BLogo and etc. (Olsen et al., 2013; Bindewald,

Schneider, & Shapiro, 2006; Workman, Yin, Corcoran, Stormo, & Benos, 2005; Li, Yang, Liang, Wang, Whiteley, Cao & Wang, 2008). Although many improved version of Sequence Logo has been developed, the original Sequence Logo remains to be the most popular graphical representation for the conserved characteristics of the sequence motifs. Table 2.1 shows the list of specialized or improved graphical representation similar to Sequence Logo.

Table 2.1: List of specialized or improved graphical representation similar to Sequence Logo.

Tools / Graphical Representation	Sequence Type	Description
BlockLogo	DNA, RNA, Protein	Visualization of protein & nucleotide fragments, continuous protein sequence motifs & discontinuous sequence motifs using the calculation of block entropy from MSA.
BLogo	DNA, RNA, Protein, Codon	Detects & displays statistically significant position-specific sequence bias with reducing background noise.
CodonLogo	DNA, RNA	Visualization of conserved regions in a MSA in the context of codons.
CorreLogo	DNA, RNA	Generates a 3D representation of properties of user-submitted RNA or DNA alignments. It calculates the mutual information of nucleotides in different positions to determine correlation and potential base pairing.
enoLOGOS	DNA, RNA	Generates Sequence Logos from a variety of input data including energy measurements, probability matrices, alignment matrices, count matrices & aligned sequences.
iceLogo	Protein	Provides a probability-based visualization by allowing users to define the reference sequence of the sample's origin.
PhylomLogo	DNA	Creates Sequence Logos for the comparison of phylogenetically distinct clades within an MSA of DNA sequences. Allows the user to visualize the global profile of whole multiple sequence alignment & visualize homologous logos or each clade simultaneously.

Table 2.1 continued

pLogo	DNA, Protein	A motif visualization in which residue heights are scaled relative to statistical significance.
PoreLogo	Protein	An automatic tool that analyses & visualize the amino acid composition of channels in transmembrane proteins.
RIILogo	RNA	Visualize inter- and intramolecular base pairing between two RNA molecules.
RNA Structure Logo	RNA	Combines the standard Sequence Logo with information about base pairing and mutual information of base pairs.
RNALogo	RNA	Creates a new graphical representation of the patterns in a multiple RNA sequence alignment with a consensus structure.
Seq2Logo	Protein	Offers the capacity to visualize amino acid sequence profiles in terms of amino acid enrichment & depletion.
Two Sample Logo	DNA, RNA, Protein	Generates a graphical representation of statistically significant position-specific differences in amino acid or nucleotide compositions between two sets of multiple aligned sequences.
ProfileGrids	Protein	A new protein sequence alignment visualization paradigm that represents an alignment as a colour-coded matrix of the residue frequency occurring at every homologous position in the aligned protein family.
GeneSlider	DNA, Protein	Helps to visualize the conservation and entropy of orthologous DNA and protein sequences by presenting them as one long sequence logo that can be zoomed in and out of, from an overview of the entire sequence down to just a few residues at a time.
Skyalign	DNA, RNA, Protein	Provides a unified framework for creating logos for both sequence alignments and profile hidden Markov models.
Sequence Logo with Glyph-based	Protein	Glyph-based enhancements, to highlight qualitatively relevant chemical insights resulting from residue substitution between sequences.

Most of the improved graphical representation mentioned above are created due to the lack of functionality or information on Sequence Logo. Table 2.2 shows the limitations found on Sequence Logo that biological scientist tried to solve by creating the improved graphical representation.

Table 2.2: Limitations found on Sequence Logo

Tools / Graphical Representation	Limitation Found on Sequence Logo
BlockLogo	Sequence logos treat each nucleotide as an individual independent position that carries no information about the relationship between the residues in the logo.
BLogo	- Every nucleotide symbol is assumed to have equal distribution with high background noise - The lack of information on the over-represented or under-represented symbols in a region of sequences.
CodonLogo	Unable to visualize codon patterns in Sequence Logo
CorreLogo	Sequence Logo does not take correlation between each position of the symbols into account.
enoLOGOS	Sequence Logo unable to show the interaction energies between the transcription factor and its binding site.
iceLogo	Sequence Logo unable to compare an experimental peptide or protein sequence set to the background of a species-specific amino acid, a position-specific background set or a background set influenced by experimental protocols.
PhylomLogo	Sequence Logo unable to visualize the whole multiple sequence alignment
RILogo	Sequence Logo unable to display RNA-RNA interaction between two RNA genes.
Seq2Logo	Focus on displaying position specific enrichment of amino acids and discarding information on the amino acid depletion
Two Sample Logo	- Motif position is mutually independent and same background distribution applied to the motif at each position. - Insensitive to sample size and unable to compare two sets of alignment

Table 2.2 continued

ProfileGrids	• Protein alignment and position with rare residues are not legible since the symbol size is small when scaled relative to the conservation position • Sequence Logo does not display any information about residues missing from alignment columns and lack of representation for gap symbol
GeneSlider	• Able to align short sequence only in which any sequence beyond 30 bases must be broken into a separate row
Skyalign	• Unable to clearly view the alignment of long sequence with the symbol bases • Lack of interactivity
Sequence Logo with Glyph-based	• Perceptual issues caused by letter size to represent the value • The arrangement of the letters on y-axis which makes the residues look more conserved than others based on the number of letters it is stacked upon

Besides these limitations, it was also found that the design of the graphical representation on Sequence Logo will cause misinterpretation of the results and cause biases in decision making while visualizing the representation. Sequence Logo was used to represent the results of an evaluation study on methods to model transcription factor sequence specificity of a Klf9 motif. It was found that the appearance of the logo might be misleading to the user compare to the actual results found from the study.

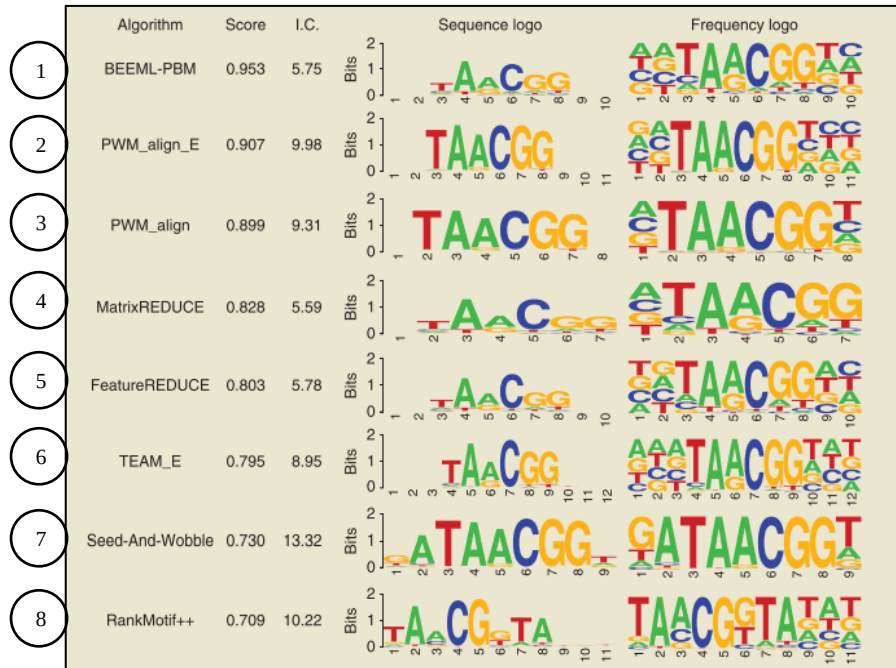


Figure 2.3: Comparison of the PWM-based algorithm evaluation results using Sequence Logo and Frequency Logo

The figure above shows the results of the evaluation study by Weirauch et.al (2013) were from the results shown by Sequence Logo, Seed-And-Wobble algorithm (7) and PWM_align algorithm (3) might be interpreted as a better predictor for the Klf9 motifs with many tall consensus sequences shown in the Sequence Logo. However, based on the evaluation results, BEEML-PBM algorithm (1) shows a better predictor of Klf9's sequence preferences compare to the two algorithms (Weirauch et al., 2013). Besides that, the importance of the initial T nucleotide in the TAACGG consensus sequence shown in the BEEML-PBM algorithm (1) might be considered as negligible based on the Sequence Logo representation. However, I was found that the nucleotide is considered highly important as other nucleotide based on the representation shown in the frequency logo on the right (Weirauch et al., 2013). Furthermore, based on the Sequence Logo, the results shown in BEEML-PBM, MatrixREDUCE, FeatureREDUCE and Team_E algorithms are nearly indistinguishable where the height of the sequence is shorter compared to the other tall

consensus Sequence Logo which is different from the results shown in the wet lab (Weirauch et al., 2013).

Moreover, previous studies have also found some limitations on the arrangement of the nucleotide in Sequence Logo and the use of Sequence Logo for comparison of the results. It was found that Sequence Logo is too focus on the information at a single position and the information shared across multiple positions are neglected (Ray et al., 2014). Additionally, limitations were found when the results are displayed in multiple separate Sequence Logo where change and inattention blindness occurs when user refocusing attention on multiple logos to understand the change (Ray et al., 2014). The process of glancing back and forth between the representations to find the difference is found to be laborious and might causes misinterpretation as it depends on the user's memory to retain the information from a representation to perform comparisons (Ray et al., 2014).

Other than that, it was found that the design of the graphical representation will also cause biases on the judgement due to the lack of information on the quality and quantity of the binding sites used in constructing a motif for the user who is non-expert on sequence analysis (Lee & Oon, 2012). The lack of indication of the height of each nucleotide will also cause confirmation biases during decision making. Moreover, the colour of the nucleotide is not standardized where users are able to choose the colour of the nucleotide they desired which causes confusion to readers during comparison (Lee & Oon, Potential Perils of Biological Sequence Visualization using Sequence Logo, 2013).

2.4 Visualization

Data visualization plays an important role in biological science and is essential to enable dissemination of scientific knowledge and sophisticated analysis of data (Sakai,

2016). The advent of technologies are able to generate, collect, manage and process a large number of data without any problem but the bottleneck in science is the ability to analyse and derive hypotheses and insights from these data that continue to increase in volume and complexity (Sakai, 2016). Traditionally, the analysis and visualization of biological data have been approached as independent problems and relatively little attention have been given to the integration and visualization of information and models. However, it is important for researchers to integrate these areas to facilitate a deeper understanding of problems in a more systematic level.

Visualization is a process of visual representation of information in the form of graphs, table, logos or symbols (Friendly, 2008). It helps to automate the description and validation of data analysis outcomes for a better understanding of the content and furthermore improve the interaction of the data representation (Friendly, 2008). In the past decades, the discovery of new biological data and algorithms has extended our ability to interrogate the biological process in many levels and the visualization methods have advanced greatly from hand-drawn pictures to computer-based visualization tool. A visualization tool can help to support three activities: the exploratory analysis, the confirmatory analysis and the presentation of data (Alexandre & Tavares, 2010). Exploratory analysis means that the user tends to discover new knowledge from the input data where they will look for signs that could suggest indications of particular relations which can use to form the basis of the hypothesis (Alexandre & Tavares, 2010). The confirmatory analysis is where a hypothesis is made by the user and therefore they will use the visual exploration of graphical representation to determine the evidence in order to accept or reject the hypothesis made (Alexandre & Tavares, 2010). Whereas for presentation, a visualization tool will create

a graphical display which can represent the structure, relationship, behaviour or characteristic of the input data (Alexandre & Tavares, 2010).

Visualization uses the abilities of the human visual system to support cognition through the use of computer-generated data representation (Bob, et al., 2002). Besides containing the design of methods and tools for browsing, formulating and displaying the predicted outcomes or complex database queries, data visualization also contains the automated description and validation of data analysis outcomes (Azuaje & Dopazo, 2005). Researchers and practitioners always rely on visualization techniques such as charts, graphs and interactive displays to solve a large range of problems. Compare to analytical or statistical techniques, they frequently prefer graphical visualization to represent the information even though analytical techniques can produce more accurate results. However, the choice of graphical representation used can either help or obstruct the user's capability to interpret the data and facilitate the analysis tasks (Nielsen, Cantor, Dubchak, Gordon, & Wang, 2010). Therefore, it is essential for researchers to choose the right visualization method to represent the information that they want to display.

The increase of new visualization methods and tools causes the visualization of information to suffer from the lack of clearly defined underlying theory that causes the tools that are produced hard to be validated and defend (Purchase, Andrienko, Jankun-Kelly, & Ward, 2008). The greatest challenge most researchers faced while developing a visual tool is the selection of suitable visual attributes in the design of graphical representation. Concerned on how the algorithms can transform data into a graphical representation especially in scientific data visualization are very important. However, when it comes to deciding the best form of representation, models and theories are always being neglected or

forgotten and are mostly resort to their own aesthetic judgment (Purchase, Andrienko, Jankun-Kelly, & Ward, 2008).

Information visualization theory helps designers to evaluate visualization with references to an established framework and predict the effect of a novel visualization method (Purchase, Andrienko, Jankun-Kelly, & Ward, 2008). In 2008, Purchase et al. discussed the role of information theory in the context of information visualization (Wang & Shen, 2011). Therefore, before visualization is further discussed we must first define the information which is the core of information visualization. According to Schneider, information is defined as a measure of the decrease in uncertainty at a receiver (Purchase, Andrienko, Jankun-Kelly, & Ward, 2008). Information theory which was developed by Claude E. Shannon defines the concept of entropy as the pillars for the Shannon Information Theory. Entropy is defined as the loss of information during transmission or a measure of disorder or randomness (Purchase, Andrienko, Jankun-Kelly, & Ward, 2008). In sequence motif visualization, Sequence Logo which is a graphical representation for displaying the consensus sequence is often displayed by following the information theory.

Besides information theory, most of the theoretical foundation we have in modern visualization theory comes from other fields such as psychology, computer graphics, statistics and etc. As mentioned earlier, information visualization can be viewed as a channel to transform the dataset to the cognitive processing user. Therefore, human factors play an important role in designing and developing suitable tools for data analysis and visualization. When we talk about visualization, the power and usefulness of visualization in representing the information of scientific data are largely due to the strength of human perception. The use of colours, symbols, shape, sizes and etc. are representing on how well the information can be symbolised in order for the user to perceive while visualizing the biological data. In

the following section, we will discuss some human factor theories that are related to visualization and how these theories will affect the interpretation and decision making the ability of the user while visualizing biological sequence motifs.

2.5 Perception

The human visual system plays an important role in visualization. It attempts to structure what we see from the graphical representation into patterns in order to make sense of the information that we want to represent (Few, 2013). When a scientific or statistical data is presented in a graphical display, the user will be asked to use their visual perception to make a quantitative judgment about the content of the display by comparing the relative sizes, colours, the orientation of the elements in the graphical representation (Few, 2013). In a true science of data visualization, the implementations of both the theories of perception and computer graphics are very important. However, visualization designers rarely take into account the human perceptual issues while designing the visualization representation of the information displayed. Graphic designers and perceptual psychologist have both long understood the issues arise but not many of them will use their wisdom in each other fields when more and more scientific data are being plotted and mode into pictures. Therefore, it is essential to implement the knowledge we get from perceptual psychologist into the graphical design of scientific data.

Human perception is always related to visualization. Hence, it is important to understand the theories of perception as it can help to improve the quality and quantity of information displayed in visualization tools. The visual features that are assigned to the data attributes must take the advantage of the strength of our visual system in order to create a visualization that is well suited to the analysis needs of the viewer. Besides that, visual

interference effect that could mask the information displayed must be avoided while designing the graphical representation.

2.5.1 Gestalt Principles of perception

Gestalt principles of perception were firstly proposed by German psychologists Max Wertheimer, Kurt Koffka and Wolfgang Köhler in 1920s. This theory was initially used in the study of psychology but the concepts of the theory have influenced many research areas such as image retrieval, visual design, graph drawing, musical studies and the design of auditory display (Goldstein, 2007). Gestalt is a German word which means shape, form or pattern. The principles state that when we perceived something, we tend to visually assemble individual objects as a whole or gestalt and not as component parts. According to Koffka (2013), he stated that “the whole is greater than the sum of its parts” shows that gestalt is actually the interplay between the parts and whole. This theory is then used to develop principles that explain how we organized individual elements into groups. The fundamental concept behind the six Gestalt principles is grouping. The six Gestalt principles of perception are the law of Pragnanz, the law of similarity, the law of good continuation, the law of proximity, the law of common fate and the law of familiarity.

Pragnanz is a German word which means a good figure. The law of pragnanz or simplicity states that ‘every stimulus pattern is seen in such a way that the resulting structure is as simple a possible’ (Koffka, 2013). It is an overriding principle which can envelop all the perceptual laws. An object is considered Pragnanz if it expresses any characteristic in a sufficiently strong way to be obvious and evocative (Alexandre & Tavares, 2010). In a more simple explanation, the human brain loves the simplicity and will tend to process simple patterns faster than complex patterns. Therefore while designing graphical representation for example graphs or charts it is easier for the user to visualize if the data are organized

systematically and logically (Padmanabhan & Choudhury, 2013). Figure 2.4 shows an example where the data in chart A is disorganized whereas in chart B is organized and the human brain tends to interpret data in chart B more easily than chart A.

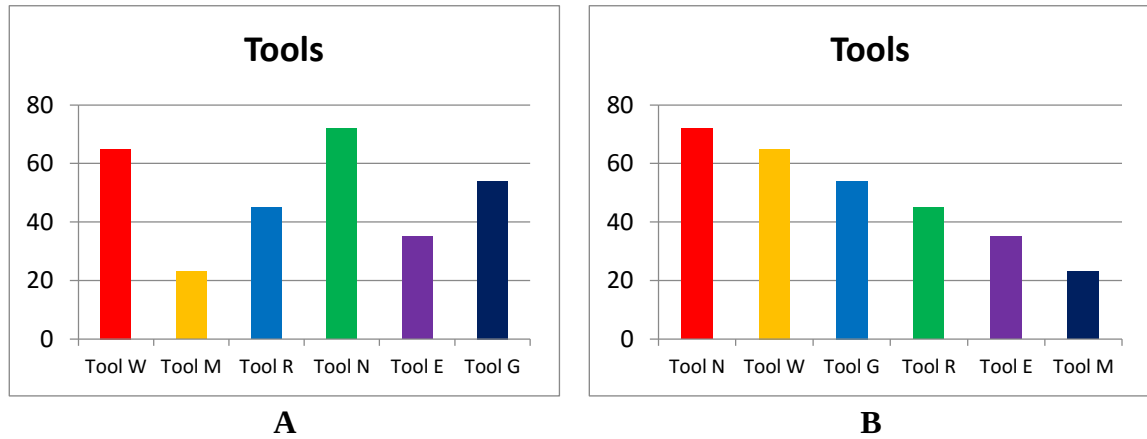


Figure 2.4: Example application of the law of Pragnanz in the arrangement of bar graphs.

Another Gestalt principle of perception is the law of similarity. The law of similarity states that ‘similar things will appear to be grouped together’ (Koofka, 2013). Our visual system will try to group elements which have attributes that are perceived as related together in the same group such as objects with similar colours, shapes, sizes, orientation, lightness and etc. (Goldstein, 2007). Although sometimes, the use of more than one colour or shapes will make the graphical display more attractive or catch more attention from the user but it may also cause cognitive overload to the user. Therefore giving similar characteristics and attributes (colour, shape, size or etc.) to the data in the same category may help to remove the additional cognitive load imposed and establish relationships between objects at the same time reinforce groupings (Padmanabhan & Choudhury, 2013). Figure 2.5 shows an example where the bars in graph B, are with the same colour which makes it easier for the user to compare and find the point of interest.

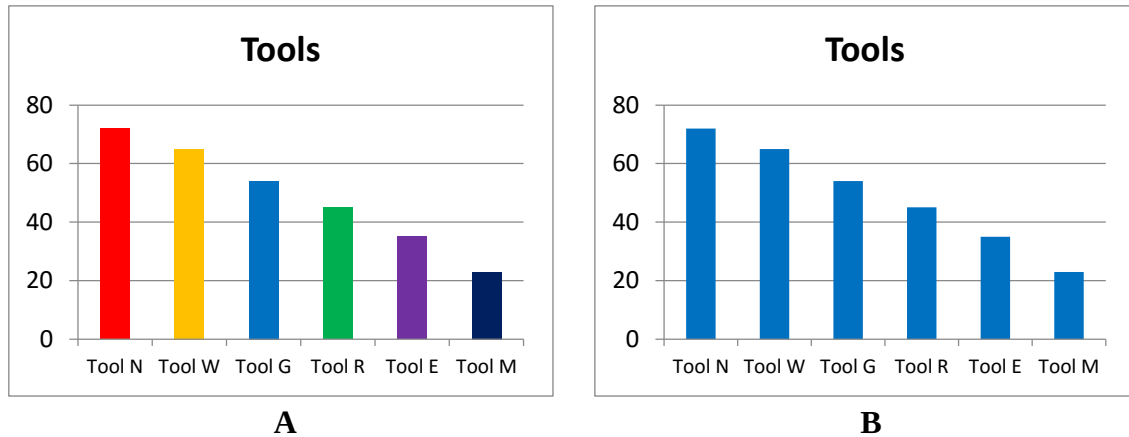


Figure 2.5: Example application of the law of similarity in the design of bar graphs.

Besides that, the law of similarity can also be applied in the concept of problems solving between a novice and an expert. For example, a group of students (novice) are asked to answer questions prepared by their teacher (expert) in physics class. The first question involved a red car with a given speed and students are asked to determine the travelling time for the red car. The second question is the calculation of force but the problem also involved a red car. Therefore, students may equate the two problems together because they both involve a red car even though they are actually two different problems. This shows that novice tends to concentrate on the superficialities in problems whereas an expert will concentrate on the deep structure of the problem. The knowledge about the thinking of novice and experts in terms of the law of similarity will help graphic designers to be more vary while choosing the type of representation for their visualization tools.

The third Gestalt principle is the law of proximity which states that ‘objects that are near to each other tend to be group together’ (Koofka, 2013). When we perceive something, we tend to make comparisons and find the similarity in the groups therefore for objects that are further apart will be seen as less related or unrelated at all. Proximity can be seen as a different grouping method compares to similarity due to the underlying spacing between objects. The relative spacing between columns or rows will dramatically affect our grouping

ability to either group the objects horizontally or vertically. The law of proximity will help to keep the data points close together by grouping them according to the established information priority. Figure 2.6 shows an example application of the law of proximity in the design of bar graphs. Users would like to compare the data between the age group of 30-40 years old in four different countries. The arrangement of the bars in graph B shows that it is easier to compare the data of the age group compare to graph A.

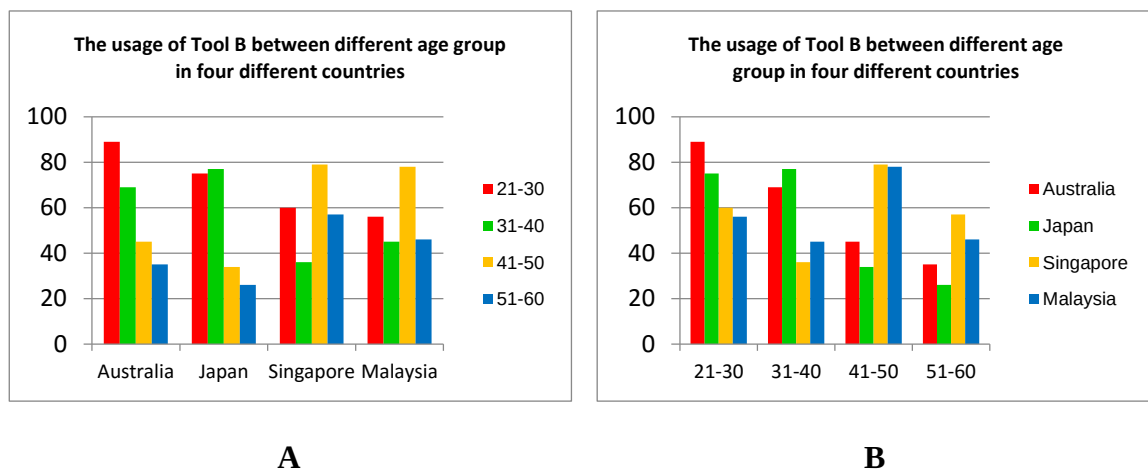


Figure 2.6: Example application of the law of proximity in the design of bar graphs.

Other than that, the fourth Gestalt principle of perception is the law of continuity. The law of continuity or good continuation states that ‘ points that, when connected, result in straight or smoothly curving lines are seen as belonging together and the lines tend to be seen in such way as to follow the smoothest path’ (Koofka, 2013). Human visual systems have a strong tendency to see shapes or lines as continuous and will tend to fill in voids with visual cues found in another place on the page (Wong, 2010). Therefore, every element on a page will affect how user perceived the graphical representation of the data in the same page. Our eyes will instinctively group things that are aligned with each other and therefore it is advised to align elements linearly to help to facilitate comparisons of different items that are in a related group. The following shows an example where the arrangement of the bars

in graph A shows abrupt directional changes whereas in graph B shows a more linear arrangement where our eyes will tend to follow the continuous path in a downward direction which makes the whole chart more readable.

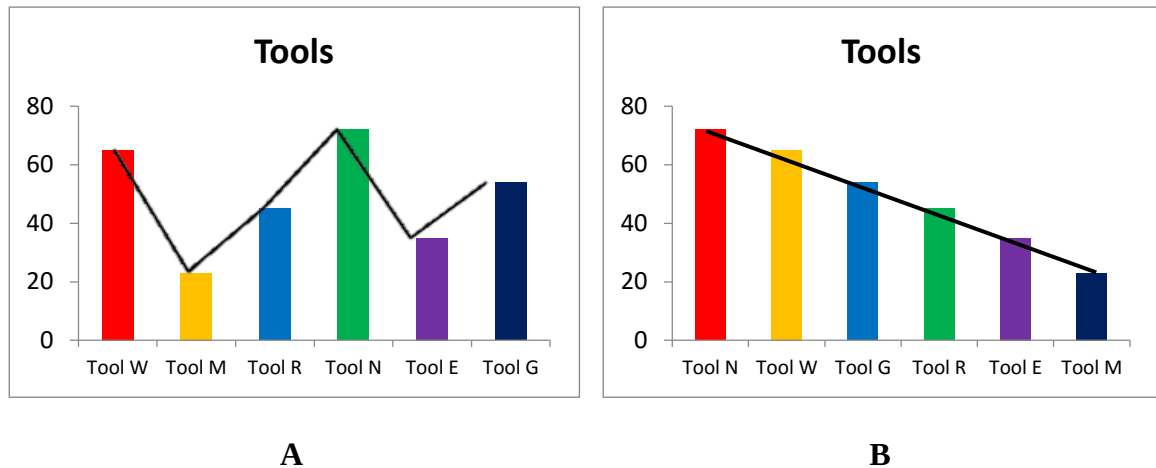


Figure 2.7: Example application of the law of continuity in the arrangement of bar graphs.

The fifth Gestalt principle is the law of common fate which states that ‘things that are moving in the same direction appear to be grouped together’ (Koofka, 2013). This law shows that if the lines or shapes are arranged or move together in the same direction (share a common fate), it will create a sense of unity which may indicate a relationship between the data whereas if the lines or shapes are arranged in different direction, they will be perceived as less related or unrelated (Padmanabhan & Choudhury, 2013). Therefore the use of direction or movement is very helpful for designers to establish the relationship between the data in the graphical representation. Figure 2.8 shows an example, the relationship of the data where the number of eggs produced by fish A and C appears to be move together in the same direction which facilitates a grouping whereas fish B shows a different direction which is perceived as different or unrelated.

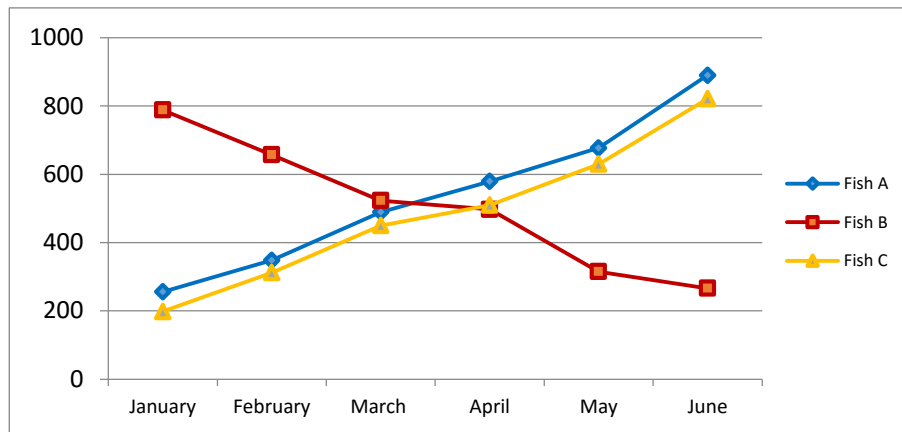


Figure 2.8: Example application of the law of common fate in the line graph.

Last but not least, the law of figure and ground states that ‘objects in a graphical display will either be perceived as figures or as a ground’ (Koofka, 2013). In this law, figure means the distinct element of focus where it will be seen as being in front of the ground whereas ground means the background where the figure placed which will be seen as unformed material that seems to extend behind the figure (Goldstein, 2007). In this law, the contrast between the figure and ground is very important as it will help the viewer to focus their attention to objects that are more important (figure) and pay less attention to the less important object (ground). The more contrast between the figure and ground, the easier for the viewer to distinguish the objects. An experiment was done by Shaun Vecera et al. (2002) to find out the factors that determine which area is figure and which area is ground. The results show that there is no left-right preference for determining figure but there is a definite preference for seeing objects lower in the display as a figure (Goldstein, 2007). Besides that, there are other four factors such as size, symmetry, orientation and meaning that will increase the chances for an area to be perceived as a figure. Figure 2.9 shows an example of the law of figure and ground where graph A which have a lower contrast between the figure and ground will cause additional cognitive load to compare to graph B which is easier to perceive.

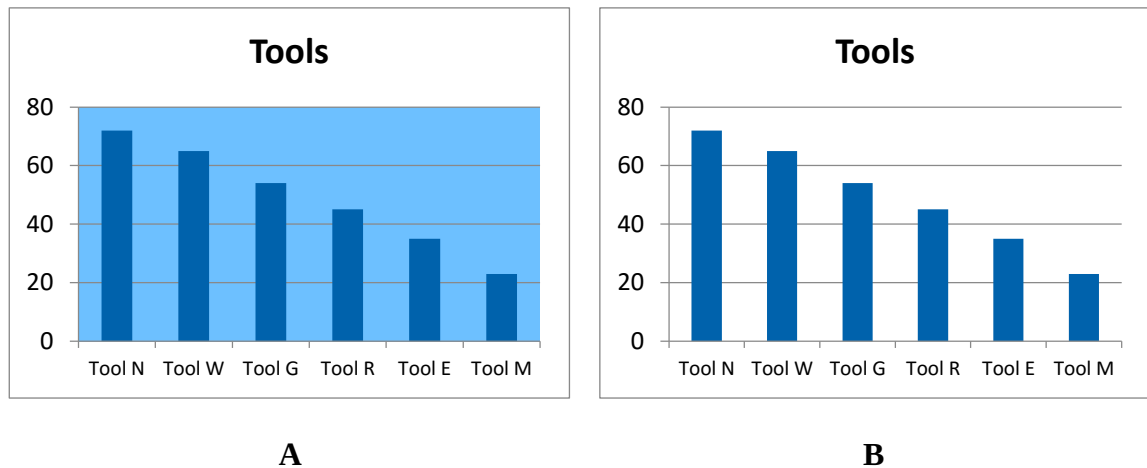


Figure 2.9: Example application of the law of figure and ground in the design of bar graphs.

Other than the six Gestalt principles proposed by the Gestalt psychologist in the early 1990s, Stephen Palmer and Palmer & Irvin have also proposed three new grouping principles to explain the perceptual organization (Goldstein, 2007). The first principle is the principle of the common region which states that ‘elements which are within the same region of space tend to be grouped together’. Another principle proposed is the principle of uniform connectedness which states that ‘a connected region of visual properties such as luminance/lightness, colour, motion, texture and etc. tend to be perceived as a single unit’ (Han, Humphreys, & Chen, 1999). Other than that, they also proposed the principle of synchrony which states that ‘the visual events that occur at the same time will be perceived as belongs to a same group’ (Goldstein, 2007). This principle is similar to the principle of common fate because both of the principles are dynamic but synchrony can occur without movement and does not need to be in the same direction.

Besides the additional three principles, some researchers from human factors in society also proposed a few principles for the design of a good user interface. The law of the focal point which is said to be the inverse of the law of similarity was proposed. This principle states that ‘distinctive objects will create a focal point in a visual presentation’

where the focal point will catch viewer's attention and attract them to follow further into the presentation (Padmanabhan & Choudhury, 2013). In order to create a focal point, designers can use colours, texture, size and etc. to capture the viewer's attention on the certain data that they want the viewer to focus on. Figure 2.10 shows an example application of the law of focal point. Viewer's attention will be drawn immediately to the highlighted bar although there are other bars which have higher values. This shows that the distinct colour of the highlighted bar creates a focal point for the following bar graph.

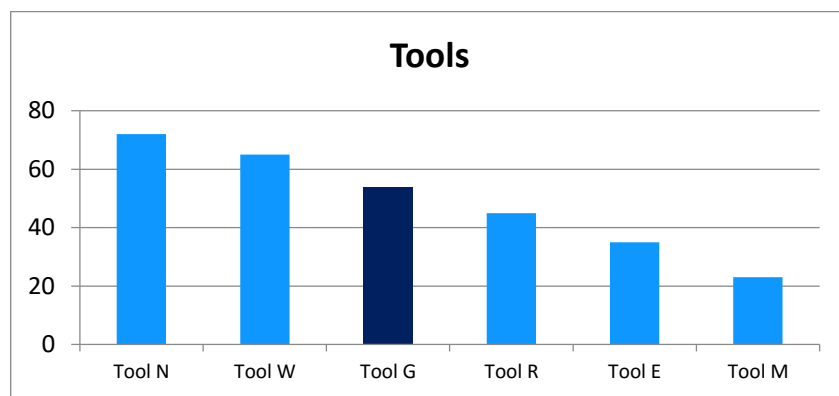


Figure 2.10: Example application of the law of focal point in the design of bar graph.

Other than that, the law of isomorphic correspondence was also proposed. This principle states that ‘human will try to interpret and respond to images based on their past experiences or knowledge about the world’ (Padmanabhan & Choudhury, 2013). For example the use of colour in the graphical representation, where from the past experience and knowledge that we learned, the red colour is always used to indicate “stop” or “danger” whereas green colour is used to indicate “go” or “nature” or “natural”. Therefore, these colours can be used while designing graphical representation which would like to express a positive negative association. Besides that, designers are also advised to use popular conventions while designing graphical representation because viewers will tend to use their past experiences to interpret the graphical display. Figure 2.11 shows an example of the law

of isomorphic correspondence where the use of red colour immediately capture viewer's attention and shows that the number of the users in that month is lower than the target size.

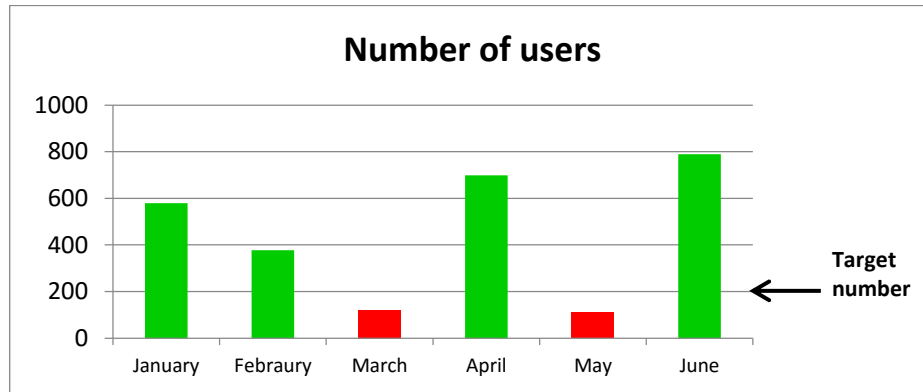


Figure 2.11: Example application of the law of isomorphic correspondence in the design of bar graph.

2.6 Attention

The human visual system contains a perceptual mechanism to help in detecting visual contours. When we visualize an object, human vision will automatically and rapidly categorize the visual images into regions and properties in parallel across the image. This phenomenon is often referred to as the preattentive processing. Preattentive visualization occurs less than 200-250 milliseconds. It allows us to detect a limited set of visual features very rapidly by low level but fast-acting visual processes. There are a few visual features that are identified as preattentive such as the orientation, length, closure, size, curvature, density, number, hue, luminance, intersections, terminators, 3D depth, flicker, the direction of motion, the velocity of motion and lighting direction (Healey & Enns, 2012). These visual attributes play an important role in capturing user's attention and organizing the data presented in a graphical representation. A few theories have been used to describe how the visual attributes help in preattentive processing and the five well-known models are the feature integration theory, textons theory, similarity theory, guided search theory and Boolean maps theory (Healey & Enns, 2012).

In 1980, Treisman and Gelade introduced the feature integration theory (FIT) of attention to describing the nature of attention and preattentive vision. The FIT states that in the preattentive stage of the perceptual processing, the features of an object are separated. When focusing attention on the object triggers the focused attention stage, the focus attention will then bind the features of the object into a coherent perception of the object (Goldstein, 2007). She also described that feature maps are encoded in parallel to the detection of feature occurs almost instantaneously in our visual system. In later work, Treisman extended the feature integration theory to explain situations in conjunction search which involve motion, depth, colour and orientation which are also preattentive visualization. Besides FIT, Julesz (1975) use the texton theory to explain the preattentive vision. He suggested that our early visual system will detect three categories of features called textons. He believes that only a difference in textons or density could be detected preattentive and no positional information about neighbouring textons is available without focus attention (Healey & Enns, 2012). Other than that, similar to Treisman, he also suggested that preattentive processing occurs in parallel and focused attention occurs in serial.

However, some researchers do not agree with the dichotomy of the serial and parallel search mode discussed by both Treisman and Julesz. Therefore, they come out with a theory called similarity theory which states that structural units that are most closely match the target template have the highest probability of accessing our visual short-term memory. The search speed is, therefore, a function of the resource allocation speed and the amount of competition for accessing our visual short-term memory (Healey & Enns, 2012). Guided search theory is then introduced by Wolfe et al. (1989) where they describe that an activation map based on both bottom-up and top-down information will be constructed during visual search. Bottom-up activation measures how different an element is from its neighbour by

following feature categorization whereas top-down activation is a user driven attempt to verify hypothesis or answer questions by searching for necessary visual information. This theory also helps to explain some cases where conjunction search can be performed preattentively because viewer driven top-down activation may help to search for conjunction target efficiently. Other than that, Huang et al. (2007) also presented a theory called Boolean map theory which divides visual search into two stages. Selection stage involves choosing a set of objects from a scene whereas the access stage determines the properties of the selected objects a viewer can apprehend (Healey & Enns, 2012). A Boolean map will help to encode the specific spatial location of the selected elements and feature labels to define the properties of the selected objects (Huang & Palsher, 2007). Besides that, some counterintuitive claims were made by using this theory such as in the search of a blue horizontal target in a sea of red horizontal and blue vertical objects. Boolean map theory states that the search for this type of combination features is difficult as it involves two Boolean map operations in series which is very unlike what stated in the FIT or guided search theory (Huang & Palsher, 2007). However, it states that the time required to search for the target object is constant and independent of the numbers of distractors.

The characteristic of preattentive vision mention above has shown how low-level visual processes can help to guide our attention in a larger scene and what visual attributes will draw our eyes in order to have focus attention to a particular object in a scene. So how should researchers choose between the different vision models? In psychophysics, the models do not compete with each other but are built on top of each other in order to address common problems and new insights over time (Healey & Enns, 2012). Therefore, it is advisable to choose the most recent model but due to the robustness of these vision models, their prediction might not be always successful for preattentive vision.

It is very important for researchers to understand visual attention while building a graphical representation as the use of proper visual features will help user to draw their attention to the areas of visualization that contains important data and correctly weight the perceptual strength of the data elements based on their encoded attribute values (Healey & Enns, 2012). One of the visual features that are most commonly used to capture our attention is the use of colour. There are some factors to consider while choosing the right colours: the distinctness, unique hues, contrast with the background, the number of colour used, field size and colour blindness (Stone, 2006). Choosing the right colour for data visualization will help to enhance and clarify the presentation whereas choosing the wrong colour will obscure, muddle and confuse the viewer (Stone, 2006). Besides colour, other visual features such as texture, size, orientation, regularity, motion flicker, direction and velocity have been continuing experiment and studied by researchers in order to produce a better guideline for data visualization.

2.7 Novice vs Expert

Perception plays an important role in visualization and different people with different level of knowledge and experience shows different perception and different way of decision making. Novice and expert are the terms that people usually used to differentiate a person's level of knowledge and experience in a domain. Expert is usually defined as a person who has achieved exceptional skills through practice and education in a domain and has an intense experience at making predictions or decision in that particular field (Phillips, Klein, & Sieck, 2004). Whereas a novice is usually a person who is new to the situation or domain with limited knowledge and experience who usually need guidance from expert in the field (Phillips, Klein, & Sieck, 2004). In order for a novice to become an expert, there is a transition on the characteristics and level of knowledge of a person and Stuart & Hubert

Dreyfus (1980) had proposed a model to display the transition of a novice to an expert in its domain.

Table 2.3: Dreyfus novice expert transition characteristics

Stages	Characteristics	Knowledge
Novice	Less experience Rigid adherence to rules taught and knowledge learned No discretionary judgement	Textbook knowledge Does not connect it to practice
Advanced beginner	Seeks for strategic and contextual knowledge Actions are based on the attributes or aspects guidelines Begins to know when rules can be broken	Working knowledge of key aspects of practice
Competent	Able to monitor own performances and cope with crowdedness Can make a conscious choice of actions Able to construct standardised and routinized procedures	Good working & background knowledge of the area of practice
Proficient	Sees situation holistically Able to differentiate the importance of a situation Increased in the use of intuition and tacit knowledge Less laboured in decision-making process	Depth understanding of discipline and area of practice
Expert	Does not need to rely on rules or guidelines Can fully adapt to or in control of the situation Intuitive grasp of the situation based on deep tacit knowledge Automatic and fluent actions	Authoritative knowledge of the discipline Deep tacit understanding across the area of practice

Stuart and Hubert Dreyfus (1980) divided the transition from a novice to expert into five developmental stages based on their skill acquisitions. Each stage shows different characteristics and level of knowledge and how their knowledge affect their action on the judgment. Many studies have been carried out to differentiate between experts and novices. Several categories were suggested such as (a) perception skills; (b) mental models; (c) sense of typicality and associations; (d) assessment of context; and (e) decision making (Klein & Militello, 2000; Anderson, 2013). In this study, we will mainly focus on the perception and decision making difference between novices and experts in visualizing a graphical display.

Different level of knowledge and experience between expert and novice shows a different ways of dealing with complexity and perception of a context (Dreyfus & Dreyfus, 1980). Table 2.4 shows the novice to expert's ways of perceiving the context and the way they can cope with complexity.

Table 2.4: Novice to the Expert perception of context

Stages	Coping with complex	Perception of context
Novice	Little or no conception in dealing with complexity	Tends to see action in isolation
Beginner	Appreciates complex situation but only able to achieve partial resolution	Sees action as a series of steps
Competent	Copes with complexity through deliberate analysis & planning	Sees actions at least partly in terms of long-term goals
Proficient	Deal complex situation holistically, more confident in decision making	Sees overall picture and how individual actions fit within
Expert	Holistic grasp of complex situations and move between intuitive and analytical approaches with ease	Sees overall picture & alternative approaches (vision of what may be possible)

The table shows that novice user will tend to perceive context in isolation form where more details information are needed for novice user to make connection between the context whereas experts tend to perceive the whole context and weight the importance of the visual information and chunk the information that is related into meaningful pattern (Norman, 2007). Therefore, different amount of information to act as cues in decision making is needed by novice and expert while perceiving a context (Brandenburg, Zander, Figge, & Beier, 2014).

2.8 Summary

This chapter discussed the function of Sequence Logo, how the Sequence Logo designs and arrange, and the current visualizing tools other than Sequence Logo that is available for user visualize the consensus sequence. This chapter also discussed the

fundamental theories that are related to human perception which are very important in designing graphical representation.

CHAPTER 3

EVALUATION STUDY ON CURRENT SEQUENCE LOGO DESIGN

3.1 Overview

This chapter discusses about the first evaluation study on the current design of the Sequence Logo. This chapter presents the research methodology used to address the first two objectives outlined in Chapter 1. It contains five major section which are the overview, research method, results of the first evaluation study, discussion and summary of this chapter. In Section 3.1 Research Method, the research design, population and sampling method, research instruments used to collect data, pilot testing and the procedures involved in data collection is discussed. Section 3.2 shows the results of the evaluation based on the hypotheses listed in Chapter 1. Section 3.3 contains the discussion for the results found and Section 3.4 summarize the chapter.

3.2 Research Method

The first evaluation study involves the evaluation of the visual attributes performance on the design of Sequence Logo based on the human factors principles in perception and visualization (Objective 1) and identify the difference between novice and experts while interpreting the result from Sequence Logo based on the Dreyfus novice expert skill acquisition model (Objective 2).

3.2.1 Research Design

This study investigates further in details on the Sequence Logo design to identify the problems found on the visual attributes representation that causes misinterpretation of the results. The problems on the visual attributes found will act as a stronger support evidence for improvement in the second evaluation study. Besides that, the way user with a different

level of experience assess the context and make a decision while interpreting the results on Sequence Logo is also evaluated to support the design of the improved Sequence Logo for the second evaluation study.

In the evaluation of the visual attributes performance, the visual attributes involved in this study are colour, the arrangement of the nucleotide, line, white space and pop up value. Each attribute are evaluated to find out which design will better help the user to interpret the information in the graphical representation. The evaluation based on user skill level in using sequence logo involves the independent variables which are the colour, arrangement, white space, line, pop up value and size whereas the dependent variable is the skill levels from novice to expert user. Novice to expert users is tested on how they assess the context in the Sequence Logo based on their perception while visualizing the logo and how they make a decision based on the information interpreted from the graphical representation.

3.2.2 Population

The populations involved are the users who had experience or background knowledge in Sequence Logo such as researcher, biologist, students and academician that had used Sequence Logo for analysing and reporting in their research. All of the participants were found through publication and universities website who were involved in the bioinformatics field or DNA sequencing field. Lecturers and postgraduate students from the Faculty of Resource Science and Technology Universiti Malaysia Sarawak were also involved in the evaluation study.

The online survey was carried out on a voluntary basis where participants were invited to participate voluntarily by email. The sample size of the respondent was set up to 100 participants however since the survey was conducted on a voluntary basis the number

of participants involved will depend on the time or availability of the participants. All the participant who received the email were given a time period of one month to answer the survey questions. This is due to some of the researcher or lecturer are on leave or are too busy to answer the survey questions. Participants who replied they are too busy but willing to forward the survey to their assistant or students are given the link to the survey to forward to them.

3.2.3 Research Instrument

3.2.3.1 Email invitation

In this survey, an invitation email is sent to researchers, students, academician and bioinformatician from all around the world to inform and request to answer the online survey. The invitation email generally consists of the purpose of the research, the time needed to complete and information related to the study. Besides that, a consent form is attached to inform the participants about the study and the confidentiality of the information shared in the survey.

3.2.3.2 Survey Questionnaire

An online survey tool called Survey Gizmo is used for the online survey. The questionnaire consists of three sections. The first section is the demographic information of the participants. The second section consists of close-ended questions based on the principles of Gestalt perception, human attention and Dreyfus novice and expert model. The close-ended questions are regarding (a) the user's perception on the colour used in current Sequence Logo visualization; (b) the user's perception on the layout design of the Sequence Logo; and (c) the sufficiency of the information provided in current Sequence Logo design. In the close-ended questions, seven scored Likert scale is used in the questions to indicate the agreement according to the respective statement. This scale consists of 7 restricted

choices of answers which are “Strongly disagree”, “Disagree”, “Slightly disagree”, “Neutral”, “Slightly agree”, “Agree” and “Strongly Agree”. Table 3.1 below shows an example of the 7 scores Likert scale.

Table 3.1: 7 score Likert scale

Strongly Disagree	Disagree	Slightly Disagree	Neutral	Slightly Agree	Agree	Strongly Agree
1	2	3	4	5	6	7

The third section consists of 4 open-ended questions regarding the problems they faced while using Sequence Logo and the recommendation on the improvement they would like to have in current Sequence Logo visualization.

3.2.3.3 Web logo

Weblogo is an online Sequence Logo generator which can help users to generate standard Sequence Logo for DNA, RNA and amino acids. In this study, all the examples of the current Sequence Logo design shown in the questionnaire are generated by using Weblogo. A sample DNA multiple sequence alignment are downloaded from JASPAR library and uploaded to Weblogo to generate the standard Sequence Logo.

3.2.4 Procedure for Data Collection

Figure 3.1 shows the procedure for data collection of the current design of Sequence

Logo:

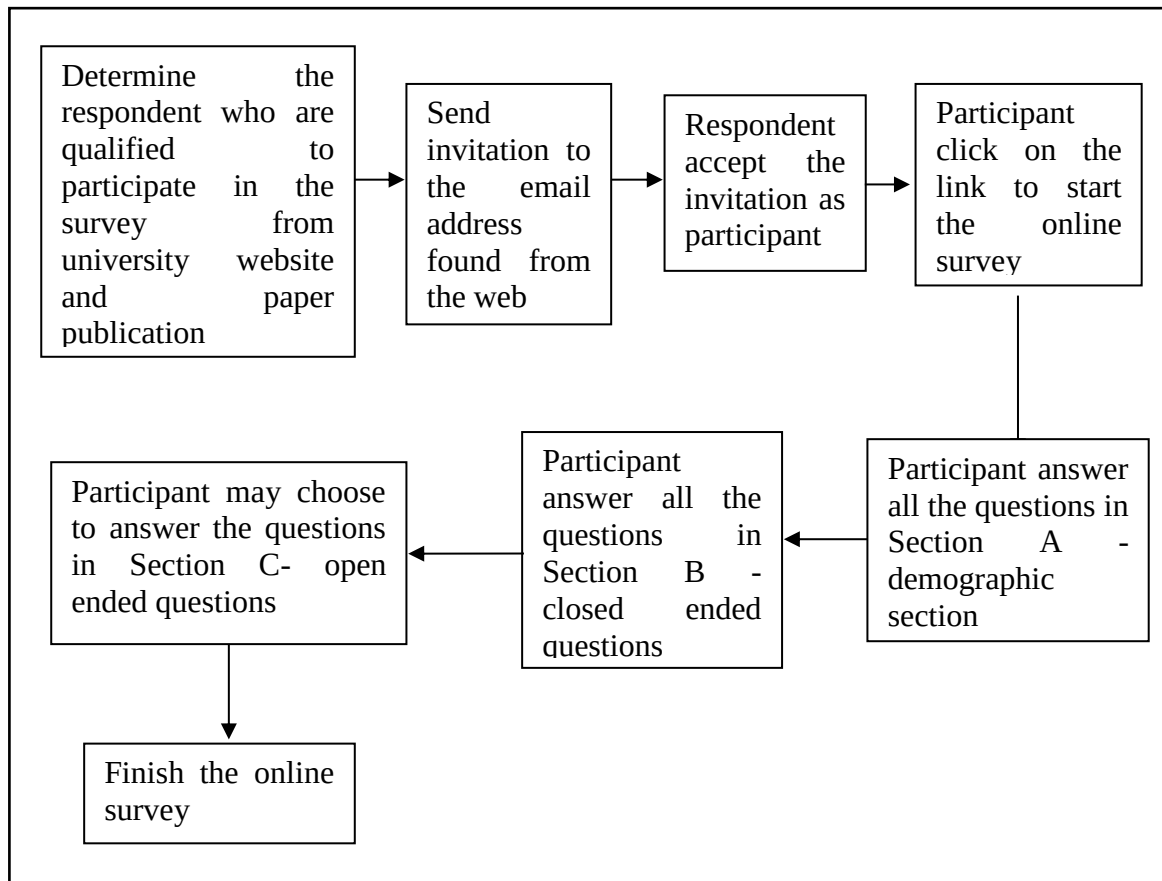


Figure 3.1: Procedure for data collection

3.2.5 Descriptive Statistics

The descriptive statistics are used to find out the mean, frequencies and percentage of the demographic information such as the gender, age, occupation, the area of expertise, skill levels, and tools used to generate Sequence Logo.

3.2.6 Inferential Statistics

There are two types of inferential statistics applied in the analysis of result of the evaluation test which is the independent sample t-test and the paired sample t-test. Independent t-test is used to compare the mean of the skill levels between the novice users

and the expert users in order to find out the difference between skill levels and the way they perceive the Sequence Logo.

Meanwhile, paired sample t-test is used to compare the Sequence Logo with a different design on colour, arrangement, the way information is presented and comparison of the results generated by Tool 1 and Tool 2 in terms of the accuracy and the quality of the Sequence Logo generated. This test will help to determine the user's perception on the design based on the visual attributes presented in the graphical representation and to test if the users are able to interpret the Sequence Logo accurately based on the design and information available.

3.3 Results

In this section, all the data obtained from the questionnaire will be analysed and evaluated. The results are divided into three parts. The demographic information closed-ended question results and the open-ended question results.

3.3.1 Demographic Information

The information gathered for a demographic section are the participant's gender, age, occupation, the area of expertise, skill levels and tools that were used to generate Sequence Logo. The results collected are shown in Table 3.2.

Table 3.2: Demographic results

Demographic	Component	Frequency (n)	Percentage (%)
Gender	Male	35	67.3
	Female	17	32.4
Age	< 30	18	34.6
	31-40	23	44.2
	41 >	11	21.1

Table 3.2: continued

Occupation	Researcher	24	46.2
	Academician	11	21.2
	Graduate Student	15	28.8
	Undergraduate Student	1	1.9
	Others	1	1.9
Area of expertise	Bioinformatics	29	55.8
	Biochemistry	4	7.7
	Sequence Analysis	4	7.7
	Genomics	6	11.5
	Molecular Biology	9	17.3

Table 3.2 shows the frequency (n) and percentage (%) of the participants in the survey based on their demographic information. There are a total of 52 participants participated in the survey. From the total of 52 participants, 67.3% of the participants are male and 32.4% are female. Most of the participants are between the age of 31 to 40 years old (44.2%) whereas 34.6% of them are below 30 years old and 21.1% are at the age of more than 41 years old. 24 (46.2%) participants work as a researcher, 11 (21.2%) participants are an academician, 15 (28.8%) are graduate students, 1 (1.9%) is an undergraduate student and 1 (1.9%) of the participant work as a bioinformatician. Most of the participants are expertise from the bioinformatics field 55.8%, 7.7% of participants involve in biochemistry, and sequence analysis field. 17.3% of the participants are from molecular biology background whereas 11.5% of the participants are expertise in the genomic field.

The result from this survey shows that the tool that is most frequently used by the participants to generate Sequence Logo is the web server tool, Weblogo. Figure 3.2 shows the tools that are used by the participants to generate Sequence Logos.

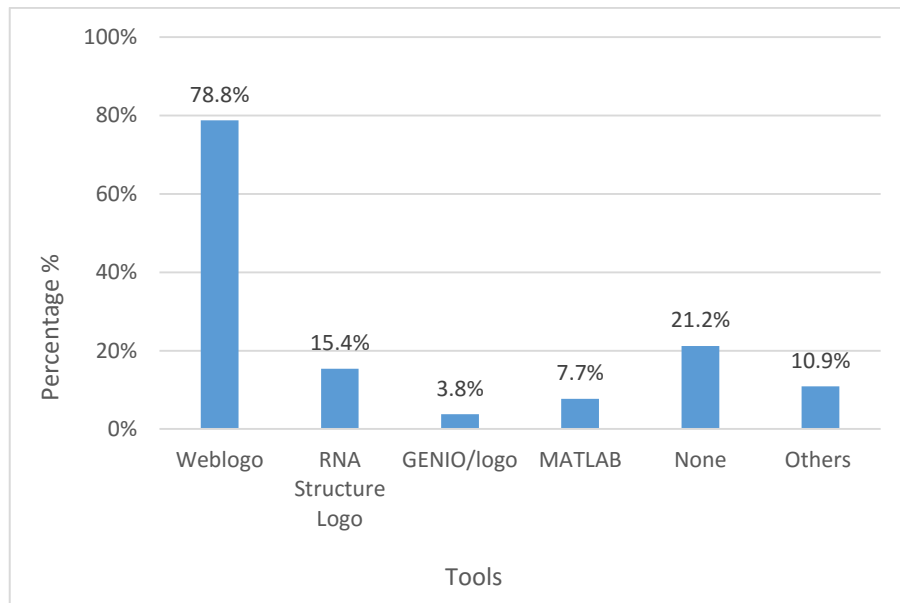


Figure 3.2: Tools to generate Sequence Logo

Figure 3.2 shows that Weblogo (78.8%) is the most popular tool used by the participants to generate Sequence Logos followed by RNA Structure Logo (15.4%), MATLAB (7.7%) and GENIO/logo (3.8%). 21.2% of the participants have not used any tools to generate Sequence Logos but only visualize it in the literature or results published by other authors. Other than the tools listed in the questionnaire, the participants also used other tools to generate Sequence Logos. 10.9% of them used other tools such as BLAST, Factorbook, HMM Logo, MEME Suite or their own script to generate Sequence Logos.

3.3.2 Closed-Ended Question Evaluation Results

The results for close-ended questions are separated into two parts, the evaluation based on user's skills in using Sequence Logo and the evaluation of the visual attributes presented on Sequence Logo.

3.3.2.1 Evaluation on the visual attributes the performance on the design of the Sequence Logo based on the human factors principles in perception and visualization (Objective 1)

In this evaluation, paired sample t-test is applied to analyse user's perception on the use of colour, line, empty space, and etc. to improve the current Sequence Logo design based on Gestalt Principle of perception. A few improvements on the current design of Sequence Logo were created based on Gestalt Principle of Similarity and Proximity where the use of colour, lines and white space is applied on the improved design to test if this improvement will help the user in visualizing and analysing the results. The following shows the paired sample t-test on the significant difference between the ease to identify the conserved nucleotide with the same colour and with the bold alphabets based on the user's perception. Figure 3.3, 3.4 and 3.5 show the original arrangement and colour of the conserved nucleotide in Sequence Logo and the recommended arrangement and colour of the conserved nucleotide of the Sequence Logo.

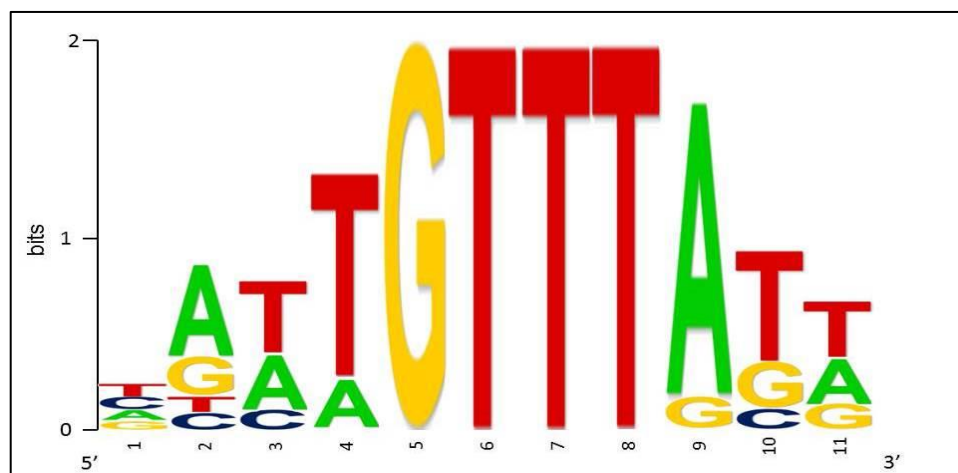


Figure 3.3: Original arrangement and colour of the conserved nucleotide in Sequence Logo.

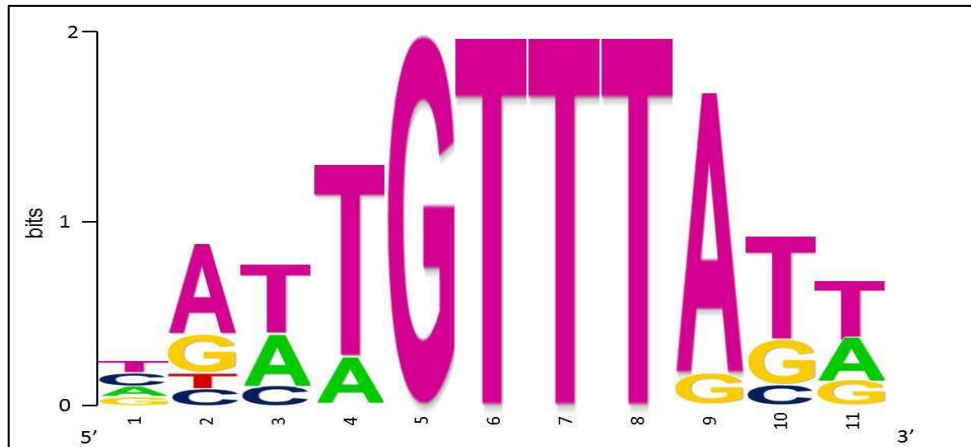


Figure 3.4: Recommended arrangement and colour of the conserved nucleotide of the Sequence Logo – the same colour of the conserved nucleotide.

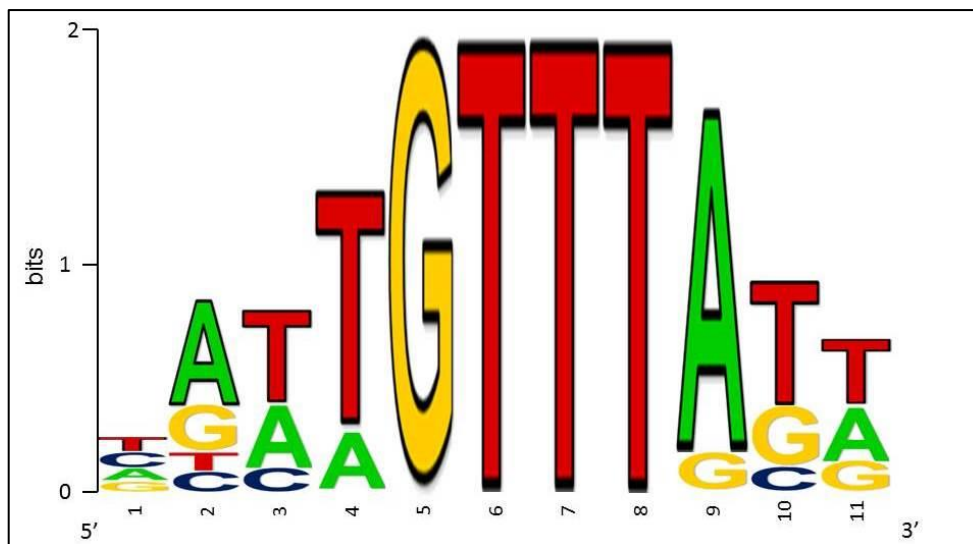


Figure 3.5: Recommended arrangement and colour of the conserved nucleotide of the Sequence Logo – Bold alphabets for the conserved nucleotide.

The null hypothesis that can be generated is:

H₀: There is no significant difference in the ease to identify the conserved nucleotide between the original arrangement and colour of the conserved nucleotide with the recommended arrangement and colour of the conserved nucleotide.

Table 3.3: Paired sample t-test for the arrangement and colour of the conserved nucleotide.

No.	Arrangement and colour of the conserved nucleotide	Mean	Std. Deviation	t	df	Sig. (2-tailed)
1	Arrange on top most position (original)	6.40	0.975	7.005	51	0.000
	Arrange on top most position with the same colour (recommended)	4.44	1.809			
2	Arrange on top most position (original)	6.40	0.975	12.494	51	0.000
	Arrange on top most position with the alphabet bold (recommended)	3.35	1.251			

With the sample size of (N=52), the result above shows that there is a significant difference between the perception of user on the original arrangement and colour of the conserved nucleotide with the recommended arrangement and colour of the conserved nucleotide where the mean and standard deviation for the original arrangement of the conserved nucleotide on the topmost position is M=6.40, SD=0.975 and the recommended arrangement of the conserved nucleotide on the topmost position with the same colour is M=4.44, SD=1.809 where the $t(51)=7.005$. The significant value for the performance is $p=0.000$ which is less than 0.05 ($p < 0.05$). Therefore, H_0 is rejected. This result shows that the arrangement of the most conserved nucleotide at the topmost position will help the user to easily identify the consensus sequence whereas the use of same colour may not preferable by the user to represent the consensus sequence.

Besides that, another recommendation where the alphabets are bold also shows a significant difference where the mean and standard deviation for the original arrangement of the conserved nucleotide on the topmost position is M=6.40, SD=0.975 and the recommended arrangement of the conserved nucleotide on the topmost position with

nucleotide alphabets bold is $M=3.35$, $SD=1.251$ where the $t(51)=12.494$. The significant value for the performance is $p=0.000$ which is less than 0.05 ($p < 0.05$). Therefore, H_0 is rejected. This result also shows that the arrangement of the most conserved nucleotide at the topmost position will help the user to easily identify the consensus sequence whereas the use of bolding alphabet may not be preferable by the user to represent the consensus sequence.

Other than the arrangement and colour for the conserved nucleotide, paired sample t-test for the recommended arrangement of the stack of the nucleotide at each position is also shown in the table below. Figure 3.6 and 3.7 show the recommended arrangement of the stack of the nucleotide at each position in the Sequence Logo.

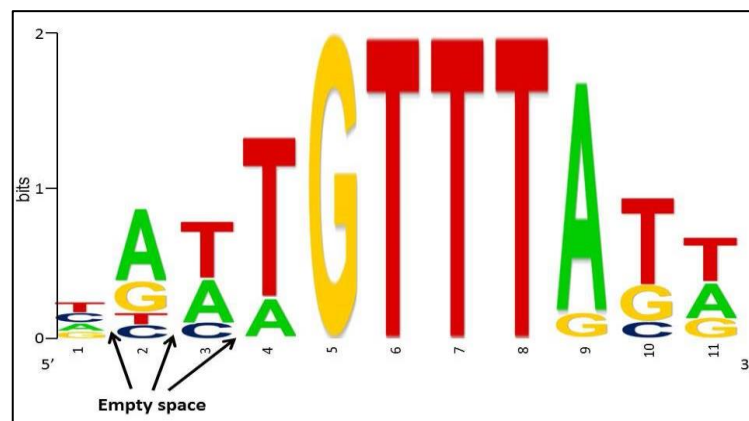


Figure 3.6: Empty spaces in the immediate adjacent position of every stack of nucleotide.

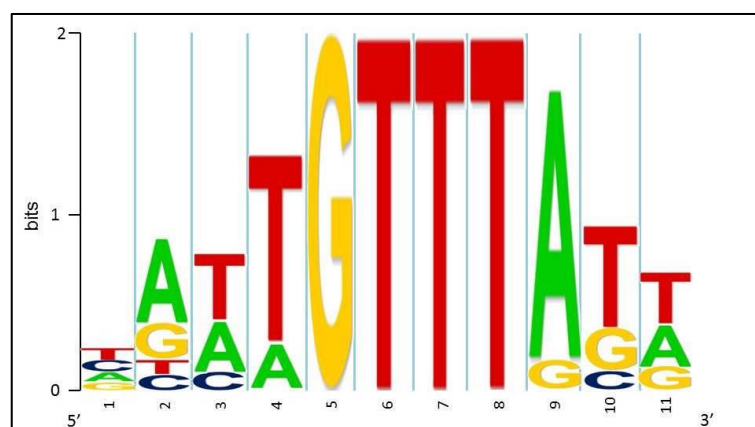


Figure 3.7: Thin line separating every stack of nucleotide adjacent to each other.

The null hypothesis that can be generated is:

H₀: There is no significant difference on the arrangement of the nucleotide by using the empty space or a thin line at the immediate adjacent position of every stack of the nucleotide to help in the grouping of the nucleotide at each position.

Table 3.4: Paired sample t-test for the arrangement of the stack of the nucleotide at each position in Sequence Logo

No.	The arrangement of the stack of the nucleotide at each position	Mean	Std. Deviation	t	df	Sig.(2-tailed)
1.	Empty spaces at the immediate adjacent position of every stack of nucleotide	4.08	1.481	3.578	50	0.001
	Thin line separating every stack of nucleotide adjacent to each other	2.82	1.646			

The results show that there is a significant difference on the arrangement of the nucleotide by using the empty space or a thin line at the immediate adjacent position of every stack of the nucleotide to help in the grouping of the nucleotide at each position. The mean and standard deviation for the arrangement of empty space at the immediate adjacent position of every stack of nucleotide in the Sequence Logo is M= 4.08 and SD=1.481 whereas the mean and standard deviation for the arrangement of a thin line to separate every stack of nucleotide adjacent to each other is M= 2.82 and SD= 1.646 where the $t(50)=3.578$. The significant value for the arrangement of the stack of the nucleotide at each position is $p=0.001$ which is less than 0.05 ($p < 0.05$) and therefore, H_0 is rejected. The result shows that the use of empty space between the nucleotide is better in grouping the stack of nucleotide than the use of thin line.

Furthermore, a paired sample the t-test is carried out to test on the ease to determine the relative conservation value of each nucleotide in the Sequence Logo. Figure 3.8 and 3.9 show the ways to determine the relative conservation values of each nucleotide in the Sequence Logo.

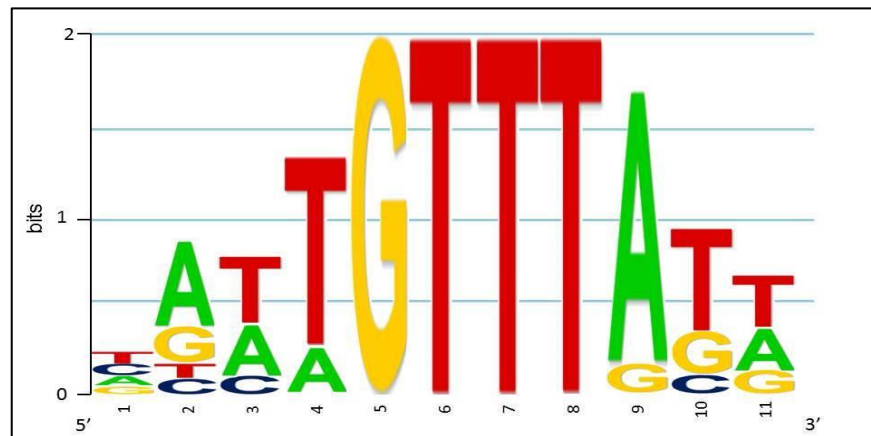


Figure 3.8: Vertical lines on the background indicating the vertical axis bit values.

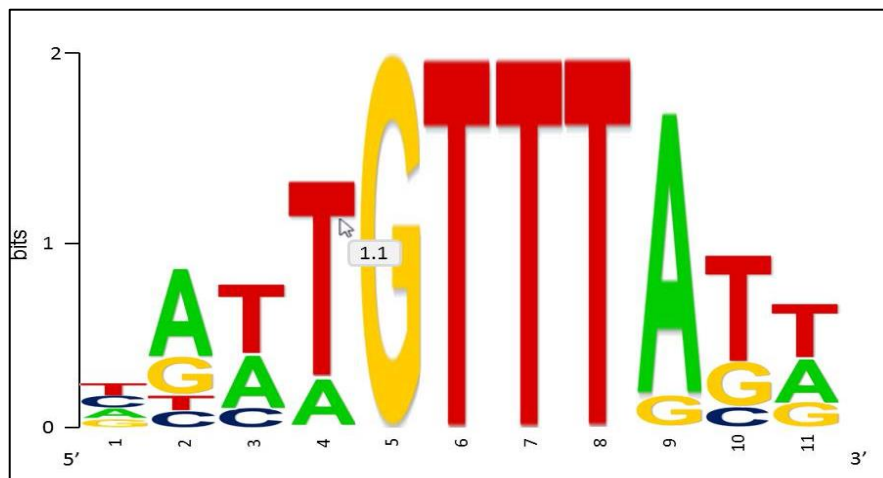


Figure 3.9: Bit values are shown when the cursor hovers over the nucleotide.

Table 3.5 shows the result of the paired sample t-test on the ease to determine the relative conservation value of each nucleotide in the Sequence Logo.

Table 3.5: Paired sample t-test on the ease to determine the relative conservation value of each nucleotide in the Sequence Logo

No.	Ways to determine the relative conservation value of each nucleotide	Mean	Std. Deviation	t	df	Sig. (2-tailed)
1.	Vertical lines on the background indicating the vertical axis bit values	3.94	2.062	-3.182	51	0.002
	Bit values are shown when the cursor hovers over the nucleotide	4.94	1.552			

The null hypothesis that can be generated is:

H₀: There is no significant difference between ease to determine the relative conservation value of each nucleotide in the Sequence Logo with the vertical lines on the background and with the bit values shown when the cursor hovers over the nucleotide.

The result in this paired sample t-test shows that there is a significant difference between ease to determine the relative conservation value of each nucleotide in the Sequence Logo with the vertical lines on the background and with the bit values shown when the cursor hovers over the nucleotide. The mean and standard deviation for the existence vertical lines on the background to indicate the vertical axis bit values is M= 3.94 and SD= 2.062 whereas the mean and standard deviation for existence of bit values shown when the cursor hovers over the nucleotide is M= 4.94 and SD= 1.552 where $t(51) = -3.182$. The significant value for the arrangement of the stack of the nucleotide at each position is $p=0.002$ which is smaller than 0.05 and therefore, H_0 is rejected. This result shows that it is easier to determine the relative conservation value by using the hover function than using the horizontal line on the background of the logo.

Besides the visual attributes, user was also test on their interpretation ability on the Sequence Logo generated by two different two to find if the user is able to accurately interpret the results in the Sequence Logo by providing an annotated motif as a reference. Table 3.6 shows the paired sample t-test on the accuracy performance, importance of T nucleotide at position 3 and the quality performance between Tool 1 & Tool 2 with the annotated motif.

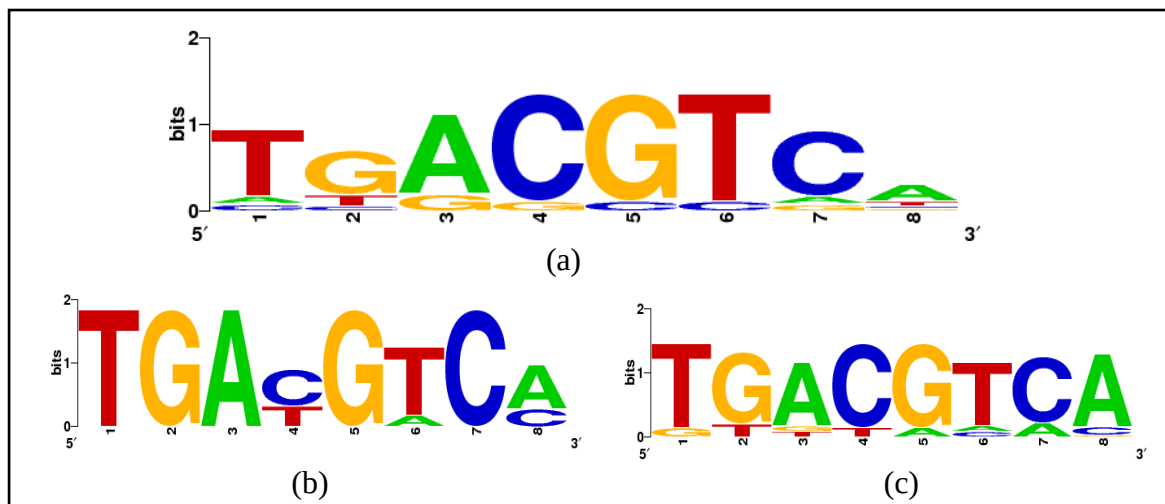


Figure 3.10: (a) Annotated Motif; (b) Motif sequence logo predicted by Tool 1; (c) Motif sequence logo predicted by Tool 2.

Table 3.6: Paired sample t-test on the performance between Tool 1 & Tool 2 with the annotated motif.

No.	Aspect	Tool	Mean	Std. Deviation	t	df	Sig. (2-tailed)
1	Accuracy	1	4.17	1.424	-4.592	51	0.000
		2	5.15	1.392			
2	Importance of nucleotide T at position 3	1	4.96	1.644	-2.213	51	0.031
		2	5.60	1.071			
3	Quality	1	4.29	1.226	-2.704	51	0.009
		2	4.92	1.118			

Table 3.6 shows the comparison of the accuracy performance, importance of T nucleotide at position 3 and the quality performance between Tool 1 & Tool 2 with the annotated motif based on user's perception while visualizing the sequence logo. The null hypothesis that can be generated is:

H₀: There is no significant difference on how user perceives and rate the Sequence Logo result between Tool 1 & Tool 2 in terms of accuracy, the importance of nucleotide T at position 3, and the quality of the result.

Based on the result, it shows that there is a significant difference on how user perceive and rate the accuracy performance of Tool 1 & Tool 2 where the mean and standard deviation for Tool 1 is M=4.17, SD=1.424 and Tool 2 is M=5.15, SD=1.392 where $t(51)=-4.592$. The significant value for the accuracy performance is $p=0.000$ which is less than 0.05 ($p < 0.05$). Therefore, H₀ is rejected. This result shows that the user will perceive the Sequence Logo results generated by Tool 2 is more accurate than Tool 1. This is further proven where the mean of Tool 2 (M=5.15) is higher than Tool 1 (M=4.17).

Besides that, the result also shows that there is a significant difference on how user perceive and rate in the importance of nucleotide T at position 3 between Tool 1 (M=4.96, SD=1.644) and Tool 2 (M=5.60, SD=1.071) where $t(51)= -2.213$. From the result, the significant value is $p=0.031$ which is less than 0.05 and therefore the H₀ is rejected. This result shows that the user will perceive the nucleotide T at position 3 in the logo generated by Tool 2 is more important than Tool 1. This is further proven where the mean of Tool 2 (M=5.60) is higher than Tool 1 (M=4.96).

In terms of the quality performance for Tool 1 and Tool 2, the result shows that there is a significant difference on how user perceive and rate the quality of the Sequence Logo

results generated Tool 1 (M=4.29, SD=1.226) and Tool 2 (M=4.92, SD=1.118) where the $t(51) = -2.704$ and the significant value is $p=0.009$ which is less than 0.05. Hence, the H_0 is rejected. This shows that user will perceive the results generated by Tool 2 as a better quality result than the results generated by Tool 1.

Besides that, participants were asked to rate the importance of the information shown in Sequence Logo. Most of the participants think that information about the position of the aligned sequence (50%), the consensus sequence (44.2%), the non-conserved nucleotide (42.3%), and the colour of the nucleotide (42.3%) are important to in Sequence Logo whereas the relative frequency (36.5%), the bits value (34.6%) and the error bars (28.8%) in Sequence Logo are slightly important to the users.

3.3.2.2 Evaluation on the difference between novice and experts while interpreting the result from Sequence Logo based on the Dreyfus novice expert skill acquisition model (Objective 2)

In the questionnaire, participants were asked to rate their level of skills (from 1 to 10) in using Sequence Logo. Table 3.7 shows the level of skills based on the Dreyfus novice expert scale (Dreyfus, Dreyfus, & Zadeh, 1987).

Table 3.7: Level of skills based on Dreyfus novice expert scale

Skills level	Stage	How context is assessed	Decision Making
1-2	Novice	Analytically	Rational
3-4	Advanced Beginner		
5-6	Competent		
7-8	Proficient	Holistically	Intuitive
9-10	Expert		

3.2.2.2.1 Assess of Context

In terms of how a context is being assessed, the stages can be separated into two different group which are then assessed by analytically for a novice to the competent stage (skills level from 1-6) and holistically for proficient to the expert stage (skills level from 7-10) (Dreyfus, Dreyfus, & Zadeh, 1987). Analytic thinking can be define as the chunking of object from the context and focusing on the attributes of the object to categorise them based on the rules whereas holistic thinking is defined as assessing the context as a whole and attend to the relationship of the object and the field with a preference on explaining such relationship (Nisbett, Peng, Incheol, & Ara, 2001). Holistic approach are experience-based knowledge and emphasis on the change, contradiction and search on multiple perspectives whereas analytical approach, uses formal logic to form inference and avoid contradiction (Nisbett, Peng, Incheol, & Ara, 2001).

An independent t-test is carried out to find the difference of how the graphical representation and information present in the Sequence Logo is being assessed between the analytic and holistic group users.

Table 3.8: Frequency & Percentage of analytic & holistic group user

Stage	Frequency	Percentage (%)	How context is assessed	Percentage (%)
Novice	5	9.62	Analytically	50
Advanced Beginner	8	15.38		
Competent	13	25		
Proficient	7	13.46	Holistically	50
Expert	19	36.54		
TOTAL	52	100	TOTAL	100

The null hypothesis that can be generated is:

H₀: There is no significant difference on how the context is being assessed in Sequence Logo between analytically (skills level 1-6) and holistically (skills level 7-10) group.

Table 3.9 shows the result of the independent t-test between the two skill groups.

Table 3.9: Independent t-test result on the assess of context

The context in Sequence Logo	Skills level	Mean	Std. Deviation	t	df	Sig. (2-tailed)
The use of different colour and the use of black and white colour to identify the nucleotide	1-6 (analytically)	6.62	0.637	-0.209	50	0.835
	7-10 (holistically)	6.65	0.689			
The use of different colour to identify the non-conserved nucleotide.	1-6 (analytically)	4.58	1.922	-2.196	50	0.033
	7-10 (holistically)	5.50	0.949			
The use of the same colour for the most conserved nucleotide.	1-6 (analytically)	5.58	1.065	5.789	50	0.000
	7-10 (holistically)	3.31	1.692			
The width between the nucleotide at each position.	1-6 (analytically)	4.65	1.522	2.716	50	0.009
	7-10 (holistically)	3.58	1.332			
Arrangement of the conserved nucleotide at the top most position.	1-6 (analytically)	6.46	0.948	0.423	50	0.674
	7-10 (holistically)	6.35	1.018			
The arrangement of the nucleotide by alphabetical order for same relative frequency.	1-6 (analytically)	4.88	1.633	-2.760	50	0.008
	7-10 (holistically)	5.88	0.864			
Use of height to represent the importance of the information content.	1-6 (analytically)	4.00	1.697	-2.965	50	0.005
	7-10 (holistically)	5.27	1.373			

Table 3.9: continued

Interpretation of the information presented in Sequence Logos.	1-6 (analytically)	3.88	1.306	-2.260	50	0.028
	7-10 (holistically)	4.88	1.840			
Sufficiency of information in Sequence Logo.	1-6 (analytically)	3.23	1.423	-4.177	50	0.000
	7-10 (holistically)	4.96	1.562			

The result in Table 3.9 shows that there is a significant difference in how the context is assessed between analytically (skills level 1-6) 50% and holistically (skills level 7-10) 50% group in terms of the colour, arrangement of the conserved nucleotide and the information that is shown in Sequence Logo.

In the context of how the non-conserved nucleotides are being assessed, it shows that there is a difference between the analytic and holistic group on the use of different colour to identify the non-conserved nucleotide. The mean and standard deviation for analytic group is $M = 4.58$, $SD = 1.922$ and for holistic group is $M = 5.50$, $SD = 0.949$ where $t(50) = -2.196$ and $p = 0.033$ ($p < 0.05$). Since the significant value is smaller than 0.05, therefore the null hypothesis is rejected. This shows that there is a significant difference on how non-conserved nucleotide in Sequence Logo being assessed in terms of colour between analytic (skills level 1-6) and holistic (skills level 7-10) group where the holistic group agree that the use of different colour for the nucleotide is very helpful in identifying the non-conserved nucleotide that present in the Sequence Logo.

Besides that, the result also shows that there is a significant difference between the two groups on the use of the same colour to represent the conserved nucleotide in Sequence Logo. The mean and standard deviation for analytic group is $M = 5.58$, $SD = 1.065$ and for holistically is $M = 3.31$, $SD = 1.692$ where $t(50) = 5.789$, $p = 0.000$ ($p < 0.05$). Since the

significant value (p) is smaller than 0.05 therefore, the null hypothesis H_0 is rejected. The result shows that analytic group user prefers the use of same colour to represent the most conserved nucleotide whereas holistic group user does not agree with the representation. This is further proven with the mean value for analytical group ($M = 5.58$) and holistically group ($M = 3.31$).

Other than colour, the result also shows a significant difference on how the context is being assessed between the analytic and holistic group in terms of the arrangement of the nucleotide at each position in the Sequence Logo. The mean and standard deviation for analytic group is $M = 4.65$, $SD = 1.522$ and for holistically is $M = 3.58$, $SD = 1.332$ where $t(50) = 2.716$, $p = 0.009$ ($p < 0.05$). Since the significant value (p) is smaller than 0.05 therefore, the null hypothesis H_0 is rejected. The analytic group user thinks that the arrangement of the nucleotide are too near to each other where there are no empty spaces at the immediate adjacent position of each stack of the nucleotide will make the Sequence Logo looks messy.

The result also shows a significant difference between the analytic and holistic group in terms of the context's arrangement when the relative frequencies of the nucleotide are identical. The mean and standard deviation for analytic group is $M = 4.88$, $SD = 1.633$ and for holistically is $M = 5.88$, $SD = 0.864$ where $t(50) = -2.760$, $p = 0.008$ ($p < 0.05$). Since the significant value (p) is smaller than 0.05 therefore, the null hypothesis H_0 is rejected. Holistic group user agrees that it will be misleading to the user when the nucleotide is arranged in alphabetical order when both of their relative frequencies are identical. This is further proven when the mean value for the holistic group ($M = 5.88$) is higher than the analytic group ($M = 4.88$).

Furthermore, the result also shows a significant difference in how the context is being assessed between the analytic and holistic group when height is used to represent the importance of the information content in Sequence Logo. The mean and standard deviation for analytic group is $M = 4.00$, $SD = 1.697$ and for holistic group is $M = 5.27$, $SD = 1.373$ where $t(50) = -2.965$, $p = 0.005$ ($p < 0.05$). Since the significant value (p) is smaller than 0.05 therefore, the null hypothesis H_0 is rejected. This shows that there is a significant difference in the way of interpretation of context where the holistic group user can easily interpret the information content of the nucleotide in the Sequence Logo by verifying their height only.

Besides that, there is a difference in the interpretation of context between the analytic group and holistic group where holistic group user can interpret the information better than analytic group user. The mean and standard deviation for analytic group is $M = 3.88$, $SD = 1.306$ and for holistically is $M = 4.88$, $SD = 1.840$ where $t(50) = -2.260$, $p = 0.028$ ($p < 0.05$). Since the significant value (p) is smaller than 0.05 therefore, the null hypothesis H_0 is rejected.

Moreover, the result also shows a significant difference between the analytic and holistic group on the amount of information shown in Sequence Logo. The mean and standard deviation for analytic group is $M = 3.23$, $SD = 1.423$ and for holistically is $M = 4.96$, $SD = 1.562$ where $t(50) = -4.177$, $p = 0.000$ ($p < 0.05$). Since the significant value (p) is smaller than 0.05 therefore, the null hypothesis H_0 is rejected. From the results, it shows that the holistic group user thinks that the information shown in Sequence Logo is sufficient them to identify and analyse the result whereas analytic group user needs more information to interpret the result shown in Sequence Logo.

Nevertheless, the results show that there is no significant difference between the use of different colour to represent the nucleotide in Sequence Logo. The mean and standard deviation for analytic group is $M = 6.62$, $SD = 0.637$ and for holistically is $M = 6.65$, $SD = 0.689$ where $t(50) = -0.209$, $p = 0.835$ ($p > 0.05$). Since the significant value (p) is greater than 0.05 therefore, the null hypothesis H_0 is accepted. This shows that there is no significant difference on the use of colour to represent the nucleotide between analytic (skills level 1-6) and holistic (skills level 7-10) group where both groups thinks that it is easier to identify the nucleotide in the Sequence Logo by using colour.

Other than that, there is no significant difference on the arrangement of the nucleotide at the topmost position to represent the conservation. The mean and standard deviation for analytic group is $M = 6.462$, $SD = 0.948$ and for holistically is $M = 6.35$, $SD = 1.018$ where $t(50) = 0.423$, $p = 0.674$ ($p > 0.05$). Since the significant value (p) is greater than 0.05 therefore, the null hypothesis H_0 is accepted. Both group users agrees that the arrangement of the conserved nucleotide at the topmost position will help the user in identifying the consensus sequence.

3.2.2.2.2 Decision Making

Besides analytic and holistic group, the level stage can also be separated into two groups in terms of the way novice and experts make a decision where novice to the proficient stage (skill level 1-8) are in a rational group whereas expert stage (skill level 9-10) are in an intuitive group. An independent t-test is carried out to test on how the graphical representation and information in Sequence Logo can help the rational thinking (skill level 1-8) and intuitive thinking (skill level 9-10) group in making a decision.

Table 3.10: Frequency & percentage of rational & intuitive group user

Stage	Frequency	Percentage (%)	Decision Making	Percentage (%)
Novice	5	9.62	Rational	63.46
Advance Beginner	8	15.38		
Competent	13	25		
Proficient	7	13.46		
Expert	19	36.54	Intuitive	36.54
TOTAL	52	100	TOTAL	100

The null hypothesis that can be generated is:

H₀: There is no significant difference in the way rational thinking group (skills level 1-8) and intuitive thinking group (skills level 9-10) perceive the information in Sequence Logo to assist them in decision making.

Table 3.11: Independent t-test for the representation of information to assist in decision making between two groups

Questions	Skills level	Mean	Std. Deviation	t	df	Sig. (2-tailed)
Vertical lines assist on the analysis of information content	1-8 (rational)	4.39	1.819	2.194	50	0.033
	9-10 (intuitive)	3.42	0.838			
The bit value is shown on hover assist on the analysis of information content	1-8 (rational)	5.53	1.124	2.129	50	0.038
	9-10 (intuitive)	4.61	1.676			
The height of the nucleotide assist on the analysis of information content	1-8 (rational)	3.74	1.661	-3.226	50	0.002
	9-10 (intuitive)	5.15	1.439			
The use of same colour for consensus sequence helps in identifying the conserved nucleotide	1-8 (rational)	5.12	1.495	4.077	50	0.000
	9-10 (intuitive)	3.26	1.727			

Table 3.11: continued

Compare few sequence logo generated by different tools to get more accurate result	1-8 (rational)	5.58	1.582	1.698	50	0.096
	9-10 (intuitive)	4.79	1.653			
Prefer Sequence Logo than Frequency Logo to represent the consensus sequence	1-8 (rational)	5.85	1.623	-0.115	50	0.909
	9-10 (intuitive)	5.89	0.875			
Interpret Sequence Logo with many tall nucleotides more superior than few tall nucleotide	1-8 (rational)	5.21	1.746	-0.948	50	0.347
	9-10 (intuitive)	5.63	1.065			
Misleading when nucleotide are arranged in alphabetical order if the relative frequency are the same	1-8 (rational)	5.15	1.544	-1.621	50	0.111
	9-10 (intuitive)	5.79	0.976			

The result shows that there is a difference in the perception of vertical lines in the background of the Sequence Logo to assist rational thinking and intuitive thinking group user on analysing the information content. The mean and standard deviation for the rational group is $M = 4.39$, $SD=1.819$ and for the intuitive group is $M=3.42$, $SD=0.838$ where $t(50)=2.194$ and the significant value $p=0.0331$ which is smaller than the α value ($p<0.05$). Therefore the null hypothesis H_0 is rejected.

Besides that, the result also shows that there is a difference in the perception of the bits value to be shown to the user between rational and intuitive thinking group where the rational thinking user prefer the bits value to be shown to help them in analysing the results in Sequence Logo. The mean and standard deviation for the rational group is $M = 5.53$, $SD=1.124$ and for the intuitive group is $M=4.61$, $SD=1.676$ where $t(50)=2.129$ and the significant value $p=0.038$ which are smaller than the α value ($p<0.05$). Therefore the null hypothesis H_0 is rejected.

Other than that, it is also shown that there is a significant difference between the rational and intuitive thinking group interpreting the information content of nucleotide by comparing the height of the nucleotide. The mean and standard deviation for the rational group is $M = 3.74$, $SD=1.661$ and for the intuitive group is $M=5.15$, $SD=1.439$ where $t(50)=-3.226$ and the significant value $p=0.002$ which are smaller than the α value ($p<0.05$). Therefore the null hypothesis H_0 is rejected. The results show that intuitive thinking group is able to analyse and make a decision by comparing the height of the nucleotide. This is further shown when the mean value for intuitive thinking group ($M=5.15$) is higher than the rational thinking group ($M=3.74$).

Moreover, the result also shows a significant difference in the perception of using the same colour to represent the consensus sequence. The mean and standard deviation for the rational group is $M = 5.12$, $SD=1.495$ and for the intuitive group is $M=3.26$, $SD=1.727$ where $t(50)=4.077$ and the significant value $p=0.000$ which is smaller than the α value ($p<0.05$). Therefore the null hypothesis H_0 is rejected. The results show that rational thinking group prefer the use of same colour to differentiate the conserved sequence from the non-conserved nucleotide. This is further shown when the mean value for rational thinking group ($M=5.12$) is higher than the intuitive thinking group ($M=3.26$).

However, the results show no significant difference between the rational and intuitive thinking group in using only one Sequence Logo results to make a decision. Based on the results both rational and intuitive group will compare a few sequence logo generated by different tools to get more accurate results. The mean and standard deviation for the rational group is $M = 5.58$, $SD=1.582$ and for the intuitive group is $M=4.79$, $SD=1.653$ where $t(50)=1.698$ and the significant value $p=0.096$ which are greater than the α value ($p>0.05$). Therefore the null hypothesis H_0 is accepted.

Besides that, there is no significant difference between the two groups in the preference of using Sequence Logo to represent the consensus sequence compare to the frequency logo. Both groups of user prefer the use of Sequence Logo compare to Frequency Logo. The mean and standard deviation for the rational group is $M = 5.85$, $SD=1.623$ and for the intuitive group is $M=5.89$, $SD=0.875$ where $t(50)=-0.115$ and the significant value $p=0.909$ which are greater than the α value ($p>0.05$). Therefore the null hypothesis H_0 is accepted.

Furthermore, there is no significant difference between the two groups where both group user will interpret Sequence Logo with many tall nucleotides as more superior than a logo with few all conserved nucleotide. The mean and standard deviation for the rational group is $M = 5.85$, $SD=1.623$ and for the intuitive group is $M=5.89$, $SD=0.875$ where $t(50)=-0.115$ and the significant value $p=0.909$ which are greater than the α value ($p>0.05$). Therefore the null hypothesis H_0 is accepted.

Both group user also show no significant difference in the perception where the arrangement of the nucleotide will be in alphabetical order if both nucleotides has the same relative frequency value. The mean and standard deviation for the rational group is $M = 5.15$, $SD=1.544$ and for the intuitive group is $M=5.79$, $SD=0.976$ where $t(50)=-1.621$ and the significant value $p=0.111$ which are greater than the α value ($p>0.05$). Therefore the null hypothesis H_0 is accepted. This shows that both group users agrees that it will be misleading to the user if the arrangement of the conserved nucleotide is arranged in alphabetical order if the relative frequency is the same.

3.3.3 Open Ended Question Results

In this section, four open-ended questions were asked to get the opinion from the participants on:

1. The purpose of using Sequence Logo.
2. The limitation faced while visualizing the graphical representation.
3. The extra information needed while interpreting the results in Sequence Logo.
4. Recommendation of improvement needed on the Sequence Logo.

From the results, 13 (25%) participants answered the open-ended questions. The following shows the results of the open-ended question.

3.3.3.1 The purpose of using Sequence Logo

The results show that 8 (61.5%) participants use Sequence Logo to find the DNA or protein sequence binding sites for probe design, distinguish closely related species and etc. 3 (23%) participants use Sequence Logo for qualitative evaluation of the predicted motif through visual comparison and motif to motif similarity comparison. 1 (7.7%) participant use the graphical representation to provide clear variable and a consensus sequence for publication on journal or research paper whereas 1 (7.7%) participant use Sequence Logo to get a quick glance to see any obvious sequence pop out of the alignment.

3.3.3.2 The limitation faced while visualizing the graphical representation

From the results, 8 (61.5%) participants said that they do not face any limitation while using Sequence Logo as a graphical representation for the consensus sequence whereas 5 (38.5%) of the participant said they faced some problem while using Sequence Logo. The following shows the limitation stated by the participants:

1. Facing great challenge while visualizing gapped motif (a variable length spacer between 2 or more motifs).
2. To focus on the consensus sequence where the more frequent nucleotide is arranged at the topmost position even if it is only a fraction more frequent than the runner-up.
3. Provide not enough information to validate the results.
4. Not suitable to visualize protein sequence as there are too many amino acids and the length of the sequence is long which is hard to use for comparison.
5. Unable to determine the number of variables of the compared sequence.

3.3.3.3 The extra information needed while interpreting the results in Sequence Logo

From the results, 9 (69.2%) does not need any extra information from the Sequence Logo whereas 4 (30.8%) participants suggest the information on the bits value to be shown in the visualization as the current visualization does not provide any information on the information content value which makes user hard to determine the information content of the nucleotide especially non-conserved nucleotide.

3.3.3.4 Recommendation of improvement needed on the Sequence Logo.

From the results, 23% of the participants provide a few suggestions of improvement on the current Sequence Logo where the bits value of the information content is important to them especially for the non-conserved nucleotide. Besides that, the participants also suggest showing the consensus sequence in string form to easily visualize the conserved nucleotide.

3.4 Discussion

The result from the survey had shown that a few limitations and problems can be found in the current visualization of Sequence Logo which will furthermore affect user's judgment or decision-making process while analysing and visualizing the consensus sequence in Sequence Logo. The limitations that were found in the current visualization of Sequence Logo are in terms of the colour, arrangement, information context and interactivity of the Sequence Logo generated by the tools. Besides that, participants involved had also given some suggestion and opinion on the improvement that they would like to have in current Sequence Logo visualization which will also be discussed in the discussion below.

From the result, it was shown that all the participants participated in this survey have the basic knowledge about Sequence Logo and most of them have the experience in using Sequence Logo. The result shows that Weblogo is the most popular tool used by the participants to generate Sequence Logo. However, 21.2% of the participants do not use any tools to generate Sequence Logo but have experience in analysing the results from Sequence Logo published in journal or research papers. Other than Weblogo, the participants also use other tools to generate Sequence Logos such as RNA Structure Logo, GENIO/Logo, Matlab, BLAST, Factorbook, HMM Logo, MEME Suite and their own script.

3.4.1 Discussion of evaluation results on the visual attributes performance in Sequence Logo

In this section, a paired sample t-test was done to find out the participant's perception of the changes of the colour, layout, arrangement and the information on the Sequence Logo visualization. In terms of the colour, it was found that most of the participants still prefer the original design of the Sequence Logo where the most conserved nucleotide are arranged at the topmost position compare to the recommended design where the most conserved

nucleotide are coloured the same or the most conserved nucleotide alphabets are made bold. Colour is the most common visual attribute used in the graphical representation to capture the viewer's attention and grouping of information (Healey & Enns, 2012). The use of the same colour on the consensus sequence helps analytic group user to easily identify and interpret the consensus sequence as it helps to differ the consensus sequence from the non-conserved nucleotide. Besides that, both analytic and holistic group agrees that the use of different colour to represent the nucleotide is very useful in Sequence Logo compare to the black and white logo as the colour can help users to identify the non-conserved nucleotide easily. However, the results also show that there is a need to standardize the colour of each nucleotide as the current tools used to generate Sequence Logo allows the user to customize the colour of the nucleotide. This function in the tool will cause confusion to the user especially during comparison of the logo if different colour is used to generate the nucleotide in Sequence Logo.

Moreover, the results also show that users prefer the recommendation on the use of empty space to separate the nucleotide compared to the use of a vertical line. Although the use of vertical line can separate the nucleotide at each position clearly the Sequence Logo will become too complicated especially when the sequence alignment are very long which cause more confusion to the users. Based on the law of figure and ground, human visual system has the ability to separate the object from the background wherein the context of Sequence Logo, the empty space or negative space between the nucleotide will serve as the background helps user to easily identify the nucleotide (object) at each position (Koofka, 2013). Additionally, the use of negative space can also help in the grouping of the nucleotide. Based on the law of proximity, the relative spacing between the nucleotide at each position

helps the user to separate the nucleotide vertically where the user can easily identify the stack of the nucleotide at each position.

Furthermore, the results also show that most of the user prefer the pop up of the bits value on the nucleotide when hover to it compare to the use of horizontal line at the background of the Sequence Logo. Although the use of horizontal line can guide user's visual system across the Sequence logo to verify the height of the sequence alignment but due to the arrangement of the nucleotide are stack on top of each other at a position, the use of horizontal line is unable to clearly indicate the information content of each nucleotide at a position compare to the pop up bits value.

These results show that the choice of selecting suitable visual attributes in a graphical representation is very important as the wrong use of visual attributes may not enhance the graphical representation but cause more confusion to the user.

Besides the performance of the visual attributes, users were tested on the ability to interpret the results in Sequence Logo. A paired sample t-test was carried out on the result obtained from the survey to test on user's perception on Sequence Logo accuracy performance, the importance of T nucleotide at position 3 and the quality performance between two different tools – MDSCAN (Tool 1), MEME (Tool 2) with the annotated motif. In this evaluation, an annotated motif is presented for the user to compare the results between the two tools. Based on the result obtained, it was shown that user is able to determine the accuracy performance, the importance of T nucleotide at position 3 and the quality performance of Tool 2 are better to Tool 1 as the result of the Sequence Logo obtained by using Tool 2 is nearer to the annotated motif.

In terms of the accuracy performance between Tool 1 and Tool 2, the result shows that the participants are able to determine which of the sequence motif generated by the tools are more accurate which is Tool 2 where the mean value for the accuracy in Tool 2 is higher than Tool 1. This can be determined by the motif pattern generated by Tool 2 which is nearer to the motif pattern in the annotated motif. Besides that, this also shows that the importance of the presence of non-conserved nucleotide in the Sequence Logo where Tool 2 is able to display more nucleotide that is present at a position compare to Tool 1. Although the function of the Sequence Logo is to display the conservation characteristic of a sequence motif, the non-conserved nucleotides are also highly important in the Sequence Logo as they also provide important information in a sequence motif. This is further proven in the second aspect which is the importance of T nucleotide at the 3rd position in the Sequence Logo generated by Tool 1 and Tool 2.

The result shows that the presence of the T nucleotide at position 3 in the Sequence Logo is very important although the sequence motif generated by Tool 1 shows higher information content and more superior than the sequence motif generated by Tool 2. Besides that, compared to the annotated motif, the sequence motif generated by Tool 2 is able to show the presence of the T nucleotide at position 3 which gives more accurate conservation detection whereas the T nucleotide at position 3 in the annotated motif is nearly indistinguishable based on the annotated motif. This shows that although the annotated motif are proven to be the most accurate result, there is still some limitation in the visualization of Sequence Logo which needs to be further improved in order to display a more accurate result for better interpretation.

Furthermore, the results from the survey also show that the participants are able to determine the quality of the Sequence Logo generated by both Tool 1 and Tool 2 where most

of the participants think that the sequence motif generated by Tool 2 has a better quality than the sequence motif generated by Tool 1. Although the sequence motif generated by Tool 1 is able to show a higher information content of the conserved sequence in fact, the sequence motif is said to be too stringent and place too much emphasis on the consensus sequence compare to the annotated motif. This has proven the evaluation study reported by Weirauch, et al., 2013 where many tall nucleotides cannot indicate the quality of the results and information such as the bits value, the number of alignment and etc. are needed to assist the user in interpreting the results. This has also proved the finding of the previous study where the current design of Sequence Logo may not suitable to evaluate the predictive ability of different biological computational tools since there is not enough information on the Sequence Logo where misinterpretation and biases may occur during the interpretation and decision-making process (Lee & Oon, 2012). Therefore, an improvement on the Sequence Logo visualization is very much needed in order to enhance the display and improve the amount of information for a better quality result for the analysis of sequence motif.

3.4.2 Discussion of evaluation results based on user's skills in using Sequence Logo

In this survey, participants are asked to rate their skill levels from 1 to 10 which indicate the novice to the expert stage. The skill levels are then used to test on the relationship between the stage of expertise and the way the user analyse the result present in the Sequence Logo.

From the results, it was found that there is a relationship between the stage of expertise in using Sequence Logo with the way the context or information in the Sequence Logo is assessed while visualizing the result. According to Dreyfus model of skill acquisition, the stages from novice to expert can be divided into five which are the novice stage, advance beginner stage, competent stage, proficient stage and the expert stage (Dreyfus, Dreyfus, &

Zadeh, 1987). Each stage shows different characteristics in the perception of a context and the way they can cope with complexity.

The result from the survey shows that novice to competent stages who usually assessed the context analytically will prefer a more distinct visualization and detailed information display in the Sequence Logo in order to guide them better in identifying the conserved nucleotide. This is shown when the analytical group prefer the dominant nucleotide at each position of the sequence to be colour the same and the arrangement of the nucleotide at each position to be separated with empty spaces in order for a neater arrangement of the result. Besides that, the analytic group user also thinks that the information present on the Sequence Logo are insufficient for them to make a judgement on the results. Analytic thinking user has the tendency to focus on the specific details of the image (Nisbett & Yuri, 2005). The lack of knowledge and experience for the novice user shows that they have little conception in dealing with complexity hence tends to see the information present in the Sequence Logo in parts. Distinct visualization such as the use of colour helps them to identify the consensus sequence easily. Hence, the details such as the relative value of nucleotide and a clearer indication of the conserved sequence are needed in order to help them in the interpretation of the result.

Holistic group user which are the proficient and expert stages, shows that they have no problems in visualizing current design of Sequence Logo where they can easily interpret the information by verifying the height of the nucleotide and the information present in the current design is sufficient for them as they do not have any problem in interpreting the information presented in the Sequence Logo. Compare to analytic perceptual, holistic perceptual processes by attending to the relationship between the object and the context due to the familiarity of the object whereas analytic perceptual processes focus on a salient object

independently from the context based on the underlying rules (Nisbett & Yuri, 2005). Holistic group user (proficient & expert stage) which have more experience in using Sequence Logo are familiar with the design and layout, does not have difficulties in visualizing and interpreting the result present in Sequence Logo as they tend to see the overall picture of the result compare to the analytic group. This can be seen when a holistic group user prefers the use of different colour to help them in identifying the non-conserved nucleotide in a glance. Besides that, the arrangement of the consensus sequence at the topmost position and the height of the nucleotide can help them to easily identify the most conserved nucleotide in the logo.

The different ways of interpreting or assessing the information in the Sequence Logo shows that there is a need for improvement in the amount of information present in the graphical representation order to accommodate the needs of the user from various skill levels.

Moreover, the result also shows that there is a relationship between the stages of expertise with the performance of decision making while using Sequence Logo to analyse the results. In terms of decision making, the novice to the expert stage can be divided into two groups where the novice to proficient stage uses rational thinking in decision making whereas the expert stage uses intuitive thinking while making a decision. Making decision while using rational thinking (novice to the proficient stage) involves finding and gathering the facts information that is related and analysing the data in details where the decision maker will attempt to optimize the knowledge they have to find the best possible alternative for the solution (Hastie & Dawes, 2010). Whereas intuitive thinking decision makers will use preconscious or a more holistic manner which based on their years of practice and experiences in making a decision (Phillips, Gary, & Sieck, 2004). This can be shown in the result where intuitive thinking user can easily interpret the information content of the

nucleotide in Sequence Logo by just verifying their height whereas rational thinking decision makers prefer the presence of vertical lines or bit values in the Sequence Logo in order to help them interpret the result more easily. Besides that, rational thinking user will compare a few sequence logo in order to get more accurate results whereas intuitive thinking user can quickly interpret the results in a glance. Decision making based on intuition are faster and are performed in a more automatic way as they are mostly based on the gut feeling (Khatri & Ng, 2000). Arguments have been made by many researchers about the use of intuitive thinking in decision making especially for the experts as they are said to be more biased and involves more feelings and emotions compare to rational decision making which uses more facts and details (Khatri & Ng, 2000). Whereas many have said that the use of intuition in decision making is better than rational decision making, but it was found that actually it was based on 'how expert you are in that field' and rational decision making are basically more useful in situations that require more logic and quantitative approach (Hastie & Dawes, 2010).

The problem of whether rational or intuitive approach is better in making a decision while visualizing the Sequence the Logo is further shown in a previous study where it was found that the current design of Sequence Logo will cause biases in the decision making processes while visualizing the result in Sequence Logo (Lee & Oon, 2012). Although the subject matter experts involve in the previous study have many years of experience in using Sequence Logo but biases still occur in the decision-making process and they have the tendency to misuse the Sequence Logo in making a judgment on the result obtained. This has further proven that actually experts (intuition decision makers) are not always good or better in interpreting the results shown in Sequence Logo compare to novice user (rational decision maker) and supporting information is very much needed in the current design of

Sequence Logo in order to help either novice or experts user in interpreting the result in Sequence Logo.

3.5 Summary

The results from this evaluation study have shown that there are some issues found on the colour, arrangement and the amount of information shown on the graphical representation where improvement on these visual attributes are needed while representing the information to assist the user with different background or skill level to interpret and analyse the sequence motif more accurately. The next chapter will discuss furthermore on the Sequence Logo design where an improved version of Sequence Logo based on the finding of this evaluation is created and evaluated by the user.

CHAPTER 4

EVALUATION STUDY ON THE IMPROVED SEQUENCE LOGO

4.1 Overview

This chapter discussed the design of an improved Sequence Logo based on the findings from the previous studies and the evaluation study on the prototype of the improved Sequence Logo.

4.2 Improvement Design on the Sequence Logo (Objective 3)

Based on the findings from the evaluation, an improved version of Sequence Logo is design and evaluated. The new graphical representation design is inspired by the Sequence Diversity Diagram (Sakai & Aerts, Sequence Diversity Diagram for comparative, 2014), Sequence Bundles (Kultys, Nicholas, Schwarz, Goldman, & King, 2014), Sequence Logo with Glyph-based (Maguire, Philippe, Sansone, & Chen, 2014) who participated in the Sequence Logo redesign contest organized by BioVis 3rd Annual Symposium on Biological Data Visualization.

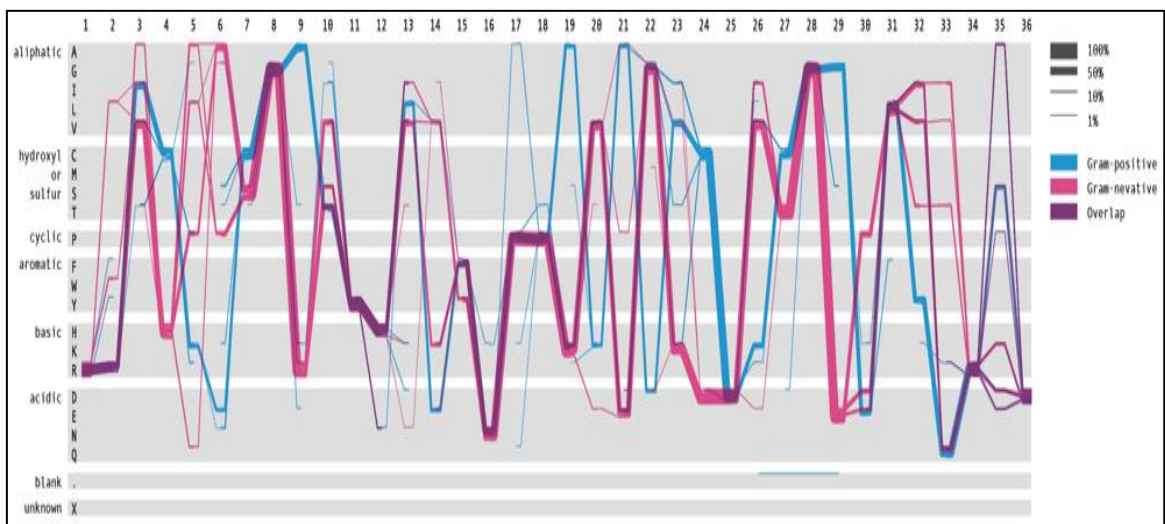


Figure 4.1: Sequence Diversity Diagram

Sequence Diversity Diagram is created to solve the problem of comparing two sets of multiple sequence alignment for protein sequence which has 14 protein families with different chemical characteristics. In order to compare two sets of multiple sequence alignments, it requires three figures for analysis: a sequence logo for across all organisms and one for each subfamily. Although each sequence logo is able to represent and summarizes the amount of information present at every position, it requires high cognitive load for the user to scan back and forth between the logos to identify key positions that are shared or differences between two subfamilies. In addition, the sequential frequency of consecutive positions is missing in the traditional sequence logo (Sakai & Aerts, 2014).

Sequence Diversity Diagram was created inspired by the Parallel Sets (ParSet) and the ProfileGrids (Sakai & Aerts, 2014). The amino acid positions are plotted horizontally on the x-axis whereas amino acid residues and their groupings are plotted on the y-axis. The grouping of amino acids is based on their structure and the general chemical characteristics are visually enhanced with horizontal gaps. A flow diagram is drawn to represent sequence alignment frequency and each sample is colour-coded. The line weight represents the relative frequency of residues at the two consecutive positions with respect to the total number of sequences. By linking two adjacent positions, it enables the user to visualize the sequential frequency and co-occurrence between two adjacent positions. By overlaying two samples, both the conserved and diverse positions are studied in a single logo (Sakai & Aerts, 2014).

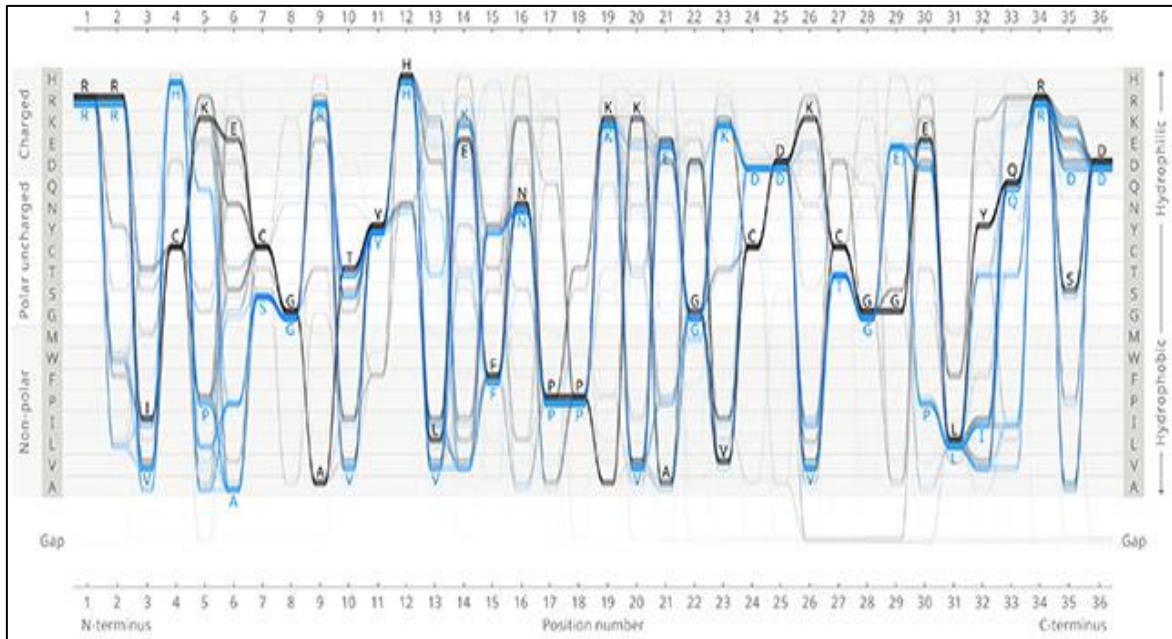


Figure 4.2: Sequence Bundles

Sequence Bundles is created to assist the discovery of hidden sequence motifs and other data features in a visualized dataset by direct manipulation and visual analysis of that data visualization itself (Kultys, Nicholas, Schwarz, Goldman, & King, 2014). In Sequence Bundles, the bundled visualization plot sequences are stacked against a Y-axis of letters arranged on a scale representing amino acid chemical characteristics. The lines' curved paths expose the conservation of residues by converging at matched positions. Their place relative to letters on the Y-axis exposes patterns in functionality. Lines representing two groups of organisms differ by colour.

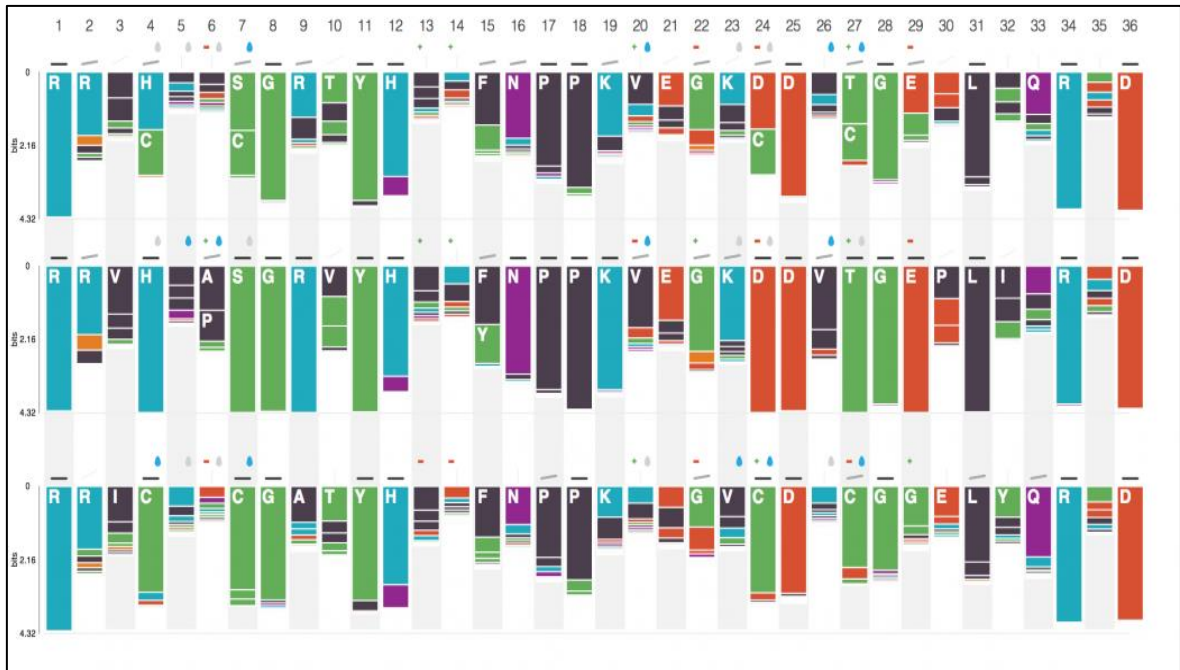


Figure 4.3: Sequence Logo with Glyph-based

Sequence Logo with Glyph-based is created to improve the current sequence logo design to address perceptual issues caused by the use of letter size to represent value and the arrangement of letters on the y-axis which can make residues look more conserved than others based on the number of letters it is stacked upon. Besides that glyphs is added to annotate positions when comparing within sets of sequence logos (Maguire, Rocca-Serra, Sansone, & Chen, 2014).

The top ‘residues’ (amino acids or nucleotides) in the Glyph-based Sequence Logo are positioned at the top or bottom of the plot so that the conserved sequence can be read more easily than in the original logo. The colours of the bars are a function of the type of amino acid, however, these are not fixed and can be changed depending on how the user wishes to visually group residues (Maguire, Rocca-Serra, Sansone, & Chen, 2014).

Based on the three redesigned Sequence Logo above, an improved Sequence Logo is designed along with the results found from the first evaluation study and Gestalt principle of perception and visualization. The following shows the Sequence logo for NHLH1 sequence alignment from JASPAR database and the prototype design of the improved Sequence Logo design for the same sequence alignment based on the underlying principles of visual attention and Gestalt principles on Perception.

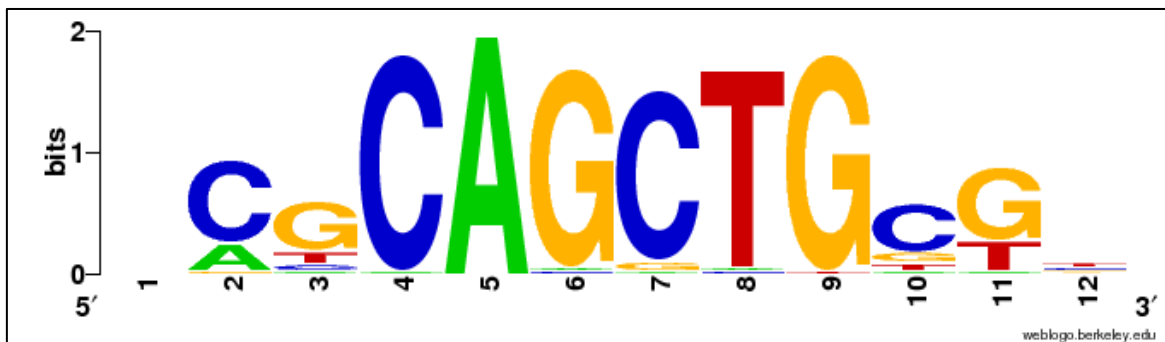


Figure 4.4: NHLH1 Sequence Logo

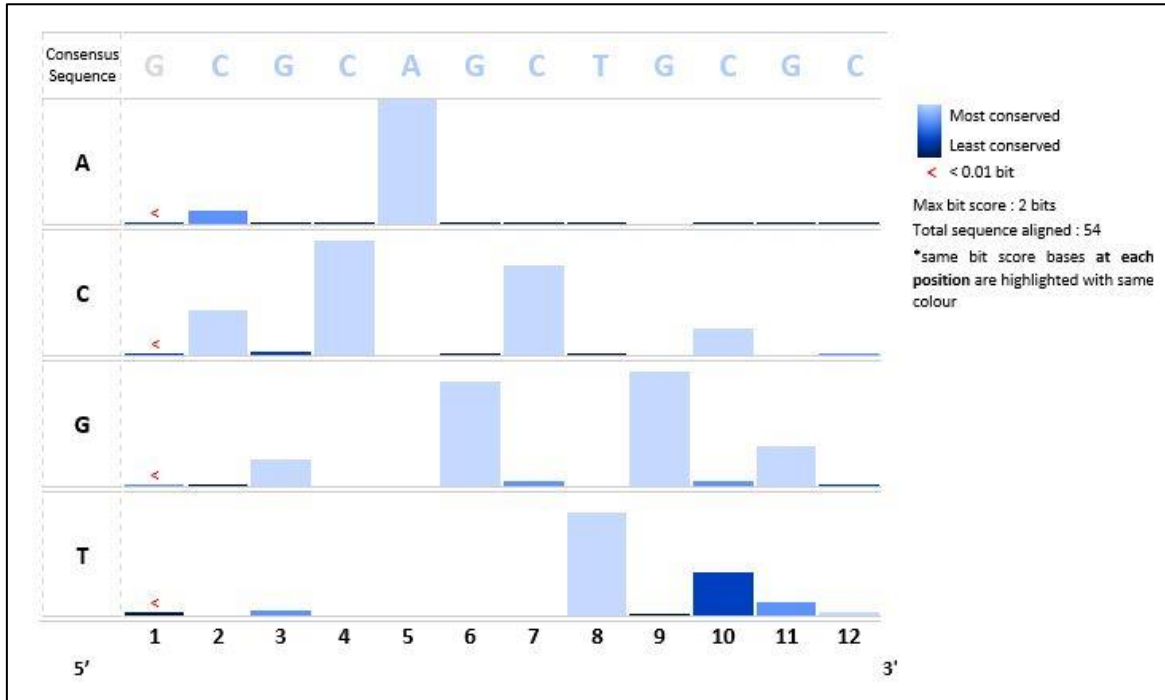


Figure 4.5: Improved Sequence Logo design

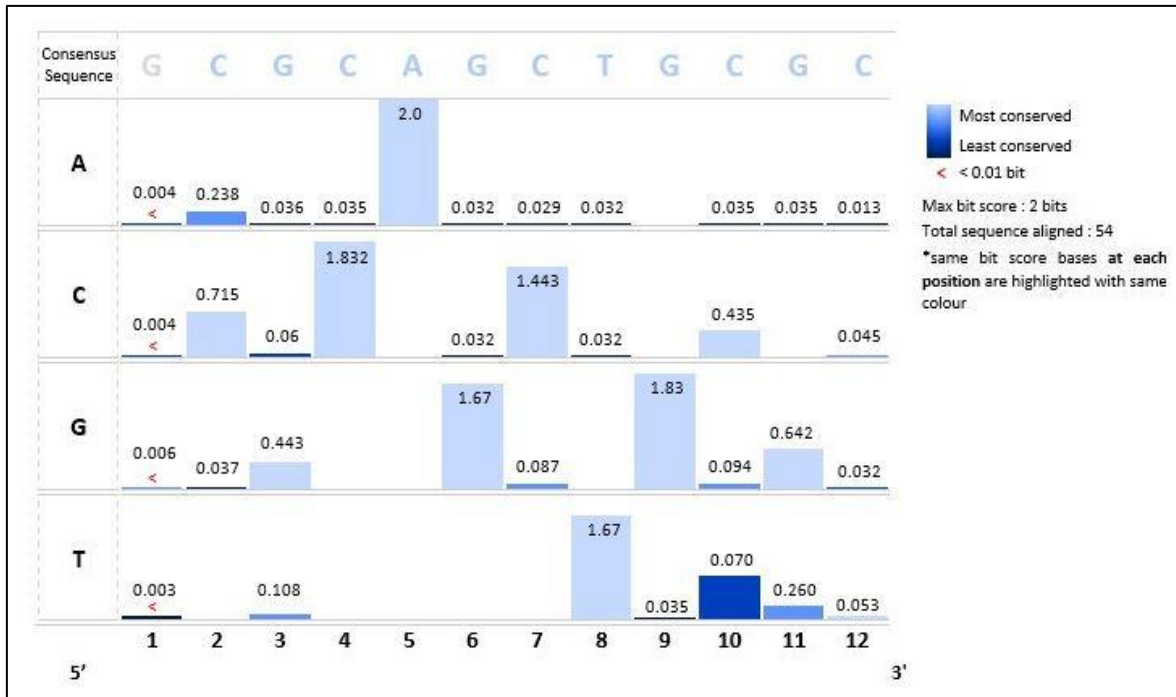


Figure 4.6: Improved Sequence Logo design with bits of value.

Inspired by the design of Sequence Logo with Glyph-based, Sequence Bundle and Sequence Diversity Diagram, the nucleotide alphabets in the improved Sequence Logo are replaced with nucleotide bars where the nucleotide A, C, G, T are arranged from top to bottom. Based on the Gestalt principles of figure and ground, the white background is used in the graphical representation to allow the user to focus on the object which is the nucleotide bar in the improved Sequence Logo. A small spacing between the nucleotide bars is used to separate and group the nucleotide at each position. The four colour (red, yellow, green, blue) used to represent the nucleotide in the original Sequence Logo are replaced with the four transition of light to deep blue colour which also represents the value of conservation from most conserved to least conserved for the nucleotide at each position. The colour is also used as the grouping element in the graphical representation to group the most conserved to least conserved nucleotide based on the law of similarity. Dotted lines are used to guide user's eyes to visualize the graphical representation from top to bottom based on the law on continuation. The consensus sequence are written clearly at the top of the graph where user

can directly read from the. Simple indication < is shown on top of the nucleotide bars that have bits value less than 0.01. User can also view the detailed bits value on top of each nucleotide bar as well as the total number of aligned sequence on the right side of the representation.

The improved Sequence Logo design is used in the evaluation study to test on user's perception of the improved graphical representation.

4.3 Research Method

The second evaluation study involves the evaluation on the user's perception on the design of the original and improved graphical representation (Objective 4) and the investigation on the difference of perception between novice and experts while interpreting the improved Sequence Logo design (Objective 5).

4.3.1 Research Design

This study evaluates the visual attributes performance in the improved design to see if the improved design are able to assist the user to perceive the information better than the original design. Besides that, the way user with a different level of experience assess the context and make a decision while interpreting the results on the improved Sequence Logo is also evaluated to find out if the design of the visual representation in the improved Sequence Logo is able to assist the user to make a better decision while perceiving the context.

In the evaluation of the visual attributes performance, the visual attributes involved in this study are colour, the arrangement of the nucleotide and the amount of information. Each attribute are evaluated to find out which design will better help the user to interpret the information in the graphical representation between the improved and the original Sequence

Logo. The evaluation based on user skill level in using sequence logo involves the independent variables which are the colour, arrangement and amount of information whereas the dependent variable is the skill levels from novice to expert user. Novice to expert users is tested on how they assess the context in the improved Sequence Logo based on their perception while visualizing the logo and how they make a decision based on the information interpreted from the graphical representation.

4.3.2 Population

The populations involved are the users who had experience or background knowledge in Sequence Logo such as researcher, biologist, students and academician that had used Sequence Logo for analysing and reporting in their research. All of the participants were found through publication and universities website who were involved in the bioinformatics field or DNA sequencing field.

The online survey was carried out on a voluntary basis where participants were invited to participate voluntarily by email. The sample size of the respondent was set up to 100 participants however since the survey was conducted on a voluntary basis the number of participants involved will depend on the time or availability of the participants. All the participant who received the email were given a time period of one month to answer the survey questions. This is due to some of the researcher or lecturer are on leave or are too busy to answer the survey questions. Participants who replied they are too busy but willing to forward the survey to their assistant or students are given the link to the survey to forward to them.

4.3.3 Research Instrument

4.3.3.1 Email invitation

In this survey, an invitation email is sent to researchers, students, academician and bioinformatician from all around the world to inform and request to answer the online survey. The invitation email generally consists of the purpose of the research, the time needed to complete and information related to the study. Besides that, a consent form is attached to inform the participants about the study and the confidentiality of the information shared in the survey.

4.3.3.2 Survey Questionnaire

An online survey tool called SoGoSurvey is used for the online survey. The questionnaire consists of two sections. The first section is the demographic information of the participants. The second section consists of close-ended questions on the user's perception of the Improved Sequence Logo design and the comparison with the original Sequence Logo design. The close-ended questions are regarding (a) the user's perception of the function of the colour in the graphical representation; (b) the user's perception on the arrangement of the nucleotide in the graphical representation; (c) the sufficiency of information provided in the representation; and (d) the perception and satisfaction of the overall design of the graphical representation. In the close-ended questions, seven scored Likert scale is used in the questions to indicate the agreement according to the respective statement. This scale consists of 7 restricted choices of answers which are "Strongly disagree", "Disagree", "Slightly disagree", "Neutral", "Slightly agree", "Agree" and "Strongly Agree". Table 4.1 shows an example of the 7 scores Likert scale.

Table 4.1: 7 score Likert scale

Strongly Disagree	Disagree	Slightly Disagree	Neutral	Slightly Agree	Agree	Strongly Agree
1	2	3	4	5	6	7

4.3.4 Procedure for Data Collection

Figure 4.7 shows the procedure for data collection of the current design of Sequence Logo:

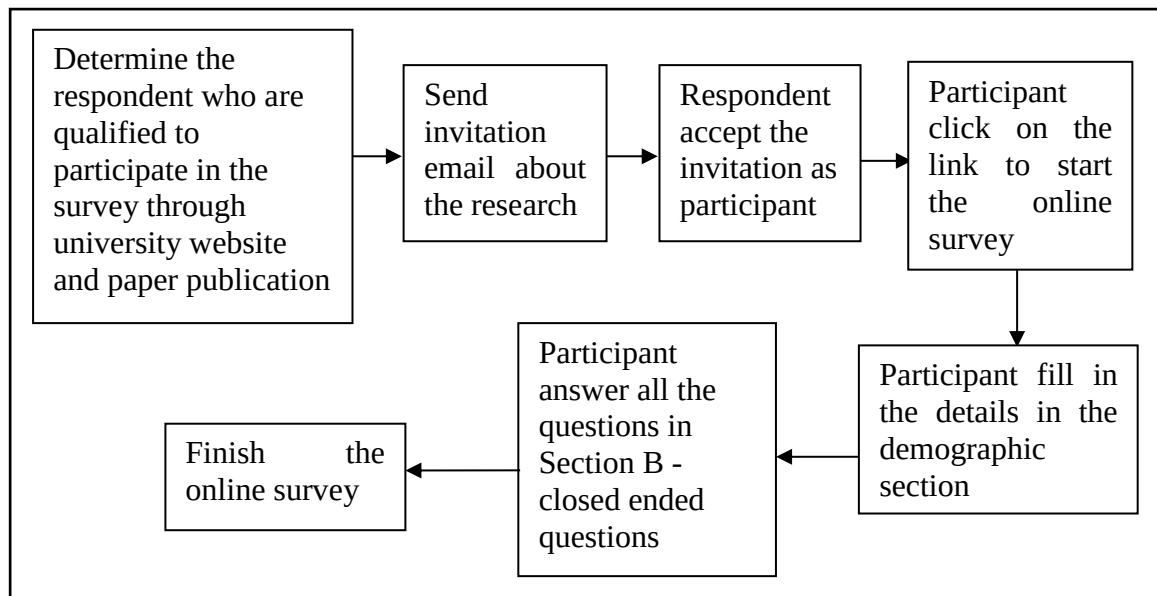


Figure 4.7: Procedure for data collection for the improved Sequence Logo evaluation

4.3.5 Descriptive Statistics

The descriptive statistics are used to find out the mean, frequencies and percentage of the demographic information such as the gender, age, occupation, the area of expertise, skill levels, tools used to generate Sequence Logo, and the purpose of using Sequence Logo.

4.3.6 Inferential Statistics

There are two types of inferential statistics applied in the analysis of result of the evaluation test which is the independent sample t-test and the paired sample t-test. Independent t-test is used to compare the mean of the skill levels between the novice users

and the expert users in order to find out the difference between skill levels and the way they perceive the original Sequence Logo and the improved design of the Sequence Logo.

Meanwhile, paired sample t-test is used to compare user's perception on the original and improved Sequence Logo design in terms of the use of colour, the arrangement of the nucleotide, the way information are presented and the overall design of both logo. This test will help to find out user's perception on the improved design if the users are able to interpret the improved Sequence Logo accurately based on the design and information available.

4.4 Results

In this section, all data obtained from the questionnaire will be analysed and evaluated. The results are divided into three parts. The descriptive results on the demographic information, the results on the independent sample t-test and the results of the paired sample t-test.

4.4.1 Demographic Information

The information gathered for a demographic section are the participant's gender, age, occupation, the area of expertise, skill levels, tools that were used to generate Sequence Logo and the purpose of using Sequence Logo. The results collected are shown in Table 4.2.

Table 4.2: Demographic results

Demographic	Component	Frequency (n)	Percentage (%)
Gender	Male	34	61.8
	Female	21	38.2
Age	< 30	22	40.0
	31-40	23	41.8
	41 >	10	18.2
Occupation	Researcher	20	36.4
	Academician	12	21.8
	Postgraduate Student	18	32.7
	Others	5	9.1

Table 4.2: continued

Area of expertise	Bioinformatics	37	67.3
	Biochemistry	2	3.6
	Genetics	11	20.0
	Molecular Biology	5	9.1
Purpose of using Sequence Logo	Graphical Representation	15	27.3
	Comparison	10	18.2
	Others	3	5.5
	All of the above	27	49.1

From the total of 55 participants, 61.8% of the participants are male and 38.2% are female. Most of the participants are between the age of 31 to 40 years old (41.8%) whereas 40% of them are below 30 years old and 18.2% are at the age of more than 41 years old. 20 (36.4%) participants work as a researcher, 12 (21.8%) participants are academicians, 18 (32.7%) are postgraduate students, and 5 (9.1%) of the participants work as a bioinformatician. Most of the participants are experts from the bioinformatics field 67.3%, 3.6% of participants involve in biochemistry, 20% in the genetics field and 9.1% of the participants are from a molecular biology background.

In this evaluation study, most of the participants use Weblogo (70.9%) to generate Sequence Logo, followed by RNA Structure Logo (10.9%). 12.7% of the participants used other tools to generate Sequence Logos such as BLAST, MEME Suite or their own script to generate Sequence Logos. Figure 4.8 shows the tools that are used by the participants to generate Sequence Logos.

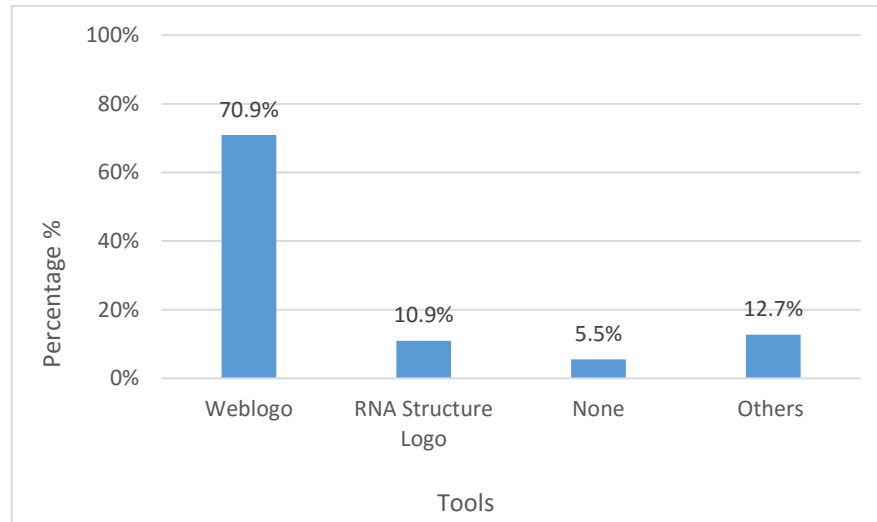


Figure 4.8: Tools to generate Sequence Logo

4.4.2 Evaluation based on user's perception of the design of the original and improved graphical representation (Objective 4)

In this evaluation, paired sample t-test is applied to analyse the perception difference between the original and improved design of the Sequence Logo on the function of colour, the arrangement of the nucleotide, the amount of information and the overall performance of the graphical representation.

4.4.2.1 Nucleotide's colour

The following shows the paired sample t-test on the function of colour in the original and improved Sequence Logo. The null hypothesis that can be generated is:

H₀: There is no significant difference in how user perceives the function of colour between the original design and improved design of Sequence Logo.

Table 4.3 shows the result of the paired sample t-test between the two graphical representations.

Table 4.3: Paired sample t-test on the function of colour in the original and improved Sequence Logo.

No.	The function of colour in the graphical representation	Sequence Logo	Mean	Std. Deviation	t	df	Sig. (2-tailed)
1	Differentiate the nucleotide A,T,G,C	Original design	6.27	0.706	40.21	54	0.000
		Improved design	1.60	0.683			
2	Identify the consensus sequence	Original design	3.75	0.93	-11.76	54	0.000
		Improved design	5.67	0.84			
3	Identify the non-conserved nucleotide	Original design	6.02	0.80	3.33	54	0.002
		Improved design	5.47	1.07			
4	Identify the level of conservation of the nucleotide	Original design	2.11	0.92	-16.80	54	0.000
		Improved design	5.20	0.91			

With the sample size of (N=55), the result above shows that there is a significant difference on the perception of how the use of colour can help in differentiating the nucleotide A, T, G, C in the original and improved Sequence Logo. The mean and standard deviation for the original Sequence Logo is M=6.27, SD=0.706 and for the improved Sequence Logo design is M=1.60, SD=0.683 where the $t(54)=40.21$. The significant value for the performance is $p=0.000$ which is less than 0.05 ($p < 0.05$). Therefore, H_0 is rejected. This result shows that user will perceive the four different colour in the original Sequence Logo as a grouping attributes for the nucleotide whereas the improved Sequence Logo does not use colour to differentiate the nucleotide.

Besides that, there is a significant difference on the perception of the function of colour in identifying the consensus sequence in the original and improved Sequence Logo. The mean and standard deviation for the original Sequence Logo is $M=3.75$, $SD=0.93$ and for the improved Sequence Logo is $M=5.67$, $SD=0.84$ where the $t(54)=-11.76$. The significant value for the performance is $p=0.000$ which is less than 0.05 ($p < 0.05$). Therefore, H_0 is rejected. This result shows that user will perceive the four transition of blue colour in the improved Sequence Logo as a grouping attributes for the conserved and non-conserved sequence whereas the original Sequence Logo does not use colour to indicate the conservation level.

Moreover, the result also shows that there is a significant difference in the perception of colour as a feature to identify the non-conserved nucleotide in the original and improved Sequence Logo. The mean and standard deviation for the original Sequence Logo is $M=6.02$, $SD=0.80$ and for the improved Sequence Logo is $M=5.47$, $SD=1.07$ where the $t(54)=3.33$. The significant value for the performance is $p=0.002$ which is less than 0.05 ($p < 0.05$). Therefore, H_0 is rejected. This result shows that the four different colour of the nucleotide can help the user to easily identify the non-conserved nucleotide in the original Sequence Logo compare to the improved Sequence Logo.

Furthermore, the result also shows that there is a significant difference in the perception of colour as a feature to identify the non-conserved nucleotide in the original and improved Sequence Logo. The mean and standard deviation for the original Sequence Logo is $M=6.02$, $SD=0.80$ and for the improved Sequence Logo is $M=5.47$, $SD=1.07$ where the $t(54)=3.33$. The significant value for the performance is $p=0.002$ which is less than 0.05 ($p < 0.05$). Therefore, H_0 is rejected. This result shows that the four different colour of the

nucleotide can help the user to easily identify the non-conserved nucleotide in the original Sequence Logo compare to the improved Sequence Logo.

Additionally, the result also shows that there is a significant difference in the perception of colour as an attribute to identify the conservation level of the nucleotide in the original and improved Sequence Logo. The mean and standard deviation for the original Sequence Logo is $M=2.11$, $SD=0.92$ and for the improved Sequence Logo is $M=5.20$, $SD=0.91$ where the $t(54)=-16.80$. The significant value for the performance is $p=0.000$ which is less than 0.05 ($p < 0.05$). Therefore, H_0 is rejected. This result shows that the use of four blue transition colour can help the user to identify the conservation level of the nucleotide compared to the colour in the original Sequence Logo.

Other than that, the preference for the use of colour in the original and improved Sequence Logo design is also analysed. The null hypothesis can be generated is:

H₀: There is no significant difference in the preference of the user in using the colour to represent the information in the original and improved Sequence Logo.

Table 4.4 shows the result of the paired sample t-test between the two graphical representations.

Table 4.4: The preference on the use of colour in the original and improved Sequence Logo

No.	Preference of the function of colour	Mean	Std. Deviation	t	df	Sig. (2-tailed)
1	Use of colour to differentiate the nucleotide.	4.38	0.952	-1.22	54	0.226
	Use of colour to identify the conservation level of the nucleotide.	4.63	1.078			

The result shows that there is no significant different on the preference of the user in using the colour to represent the information in the original and improved Sequence Logo. The mean and standard deviation for the original Sequence Logo is $M=4.38$, $SD=0.952$ and for the improved Sequence Logo is $M=4.63$, $SD=1.078$ where the $t(54)=-1.22$. The significant value for the performance is $p=0.226$ which is greater than 0.05 ($p > 0.05$). Therefore, H_0 is accepted.

4.4.2.2 Nucleotide's Arrangement

Besides the perception of colour, the perception of the arrangement of the nucleotide is evaluated. The following shows the paired sample t-test on the perception of the arrangement of a nucleotide in the original and improved Sequence Logo. The null hypothesis that can be generated is:

H₀: There is no significant difference on how the arrangement of the nucleotide in the original design and improved design of Sequence Logo helps the user to identify the consensus sequence in the representation.

Table 4.5 table shows the result of the paired sample t-test between the two graphical representations.

Table 4.5: Paired sample t-test on the arrangement of the nucleotide in the original and improved Sequence Logo.

No.	Arrangement of nucleotide	Sequence Logo	Mean	Std. Deviation	t	df	Sig. (2-tailed)
1	Easily identify the height of each nucleotide at each position	Original design	3.04	1.305	-12.97	54	0.000
		Improved design	5.67	0.668			

Table 4.5: continued

2	Easily identify the total height of the nucleotide stack at each position	Original design	5.72	0.592	26.19	54	0.000
		Improved design	2.69	0.605			
3	Able to easily compare the height of a nucleotide present at each position	Original design	2.45	0.603	-	54	0.000
		Improved design	5.67	0.668	26.63		
4	Identify which nucleotide present and not present at a position easily	Original design	4.76	1.053	-	54	0.016
		Improved design	5.20	0.704	2.486		
5	Easily comprehend the nucleotide at position 1 as all nucleotide are present with similar conservation value	Original design	5.05	1.325	-4.45	54	0.000
		Improved design	5.76	0.469			
6	Easily identify the consensus sequence.	Original design	5.65	0.517	6.524	54	0.000
		Improved design	4.92	0.604			

The result shows that there is a significant difference on how the arrangement of the nucleotide in the original design and improved design of Sequence Logo helps the user to identify the height of each nucleotide at each position. The mean and standard deviation for the original Sequence Logo is $M=3.04$, $SD=1.305$ and for the improved Sequence Logo design is $M=5.67$, $SD=0.668$ where the $t(54)=-12.97$. The significant value is $p=0.000$ which is less than 0.05 ($p < 0.05$). Therefore, H_0 is rejected. The result shows that the design of the Improved Sequence Logo can help the user to identify the height of each nucleotide better than the original Sequence Logo design. This is further proven when the mean value for the improved Sequence Logo ($M=5.67$) is higher than the original representation ($M=3.04$).

Moreover, the results also shows a significant different on the how the arrangement of the nucleotide assist user to visualize the total height of the nucleotide stack in the original and improved Sequence Logo. The mean and standard deviation for the original Sequence Logo is $M=5.72$, $SD=0.592$ and for the improved Sequence Logo is $M=2.69$, $SD=0.605$ where the $t(54)=26.19$. The significant value for the performance is $p=0.000$ which is less than 0.05 ($p < 0.05$). Therefore, H_0 is rejected. This result shows that the arrangement of the nucleotide in the original Sequence Logo is better for visualizing the total hthe eight of the nucleotide stack compares to the improved Sequence Logo.

Additionally, the result shows a significant difference on how the arrangement can help the user to compare the height of each nucleotide present in the alignment. The mean and standard deviation for the original Sequence Logo is $M=2.45$, $SD=0.603$ and for the improved Sequence Logo is $M=5.67$, $SD=0.668$ where the $t(54)=-26.63$. The significant value for the performance is $p=0.000$ which is less than 0.05 ($p < 0.05$). Therefore, H_0 is rejected. This shows that the arrangement of the nucleotide in the improved Sequence Logo can help the user to identify and compare the conservation level between the same nucleotide at the different position better than the original Sequence Logo.

Furthermore, the result also shows that there is a significant difference on how the arrangement of the nucleotide can help the user to identify which nucleotide present or not present at a position. The mean and standard deviation for the original Sequence Logo is $M=4.76$, $SD=1.053$ and for the improved Sequence Logo is $M=5.20$, $SD=0.704$ where the $t(54)=-2.486$. The significant value for the performance is $p=0.016$ which is less than 0.05 ($p < 0.05$). Therefore, H_0 is rejected. This result shows that the arrangement of improved Sequence Logo can help the user to easily identify which nucleotide present or not present at a position easily.

Other than that, the result also shows that there is a significant difference on how the arrangement of the nucleotide can help the user to interpret the nucleotide at position 1 as all nucleotide present with similar conservation value. The mean and standard deviation for the original Sequence Logo is $M=5.05$, $SD=1.325$ and for the improved Sequence Logo is $M=5.76$, $SD=0.469$ where the $t(54)=-4.45$. The significant value for the performance is $p=0.000$ which is less than 0.05 ($p < 0.05$). Therefore, H_0 is rejected. The result shows that the improved Sequence Logo can help the user to easily comprehend as all the nucleotide present has similar conservation value at position 1 compare to the original Sequence Logo.

In addition, the result also shows a significant difference on how the arrangement of the nucleotide can help the user to identify the consensus sequence in the original and improved Sequence Logo. The mean and standard deviation for the original Sequence Logo is $M=5.65$, $SD=0.517$ and for the improved Sequence Logo is $M=4.92$, $SD=0.604$ where the $t(54)=6.524$. The significant value for the performance is $p=0.000$ which is less than 0.05 ($p < 0.05$). Therefore, H_0 is rejected. The result shows that the arrangement of the consensus sequence in the original Sequence Logo can help the user to identify the consensus sequence better than the improved Sequence Logo.

4.4.2.3 Amount of Information

Furthermore, paired sample t-test is also used to analyse user's perception of the information provided on the original and improved Sequence Logo to assist in analysing the results. The null hypothesis that can be generated is:

H₀: There is no significant difference on the information present in the original and improved Sequence Logo to help the user in analysing the results.

Table 4.6 shows the result of the paired sample t-test between the two graphical representations.

Table 4.6: Paired sample t-test on the amount of information in the original and improved Sequence Logo.

No.	Information on the graphical representation	Sequence Logo	Mean	Std. Deviation	t	df	Sig. (2-tailed)
1	Information provided are adequate	Original design	3.71	1.227	-1.066	54	0.291
		Improved design	3.95	1.113			
2	Easily identify how many sequence aligned	Original design	2.29	0.685	-27.65	54	0.000
		Improved design	5.71	0.567			
3	Comprehend the logo as no nucleotide present at position 1	Original design	4.09	1.405	9.443	54	0.000
		Improved design	2.16	0.660			

The results from the paired sample t-test show that there is a significant difference the information shown in the original and improved Sequence Logo design to identify the number of sequences aligned to generate the Sequence Logo. The mean and standard deviation for the original Sequence Logo is M=2.29, SD=0.685 and for the improved Sequence Logo is M=5.71, SD=0.567 where the $t(54)=-27.65$. The significant value for the performance is $p=0.000$ which is less than 0.05 ($p < 0.05$). Therefore, H_0 is rejected. This shows that the information shown in the improved Sequence Logo can help the user to easily identify the amount of sequence aligned to generate the logo whereas the original design does not provide the information.

Besides that, the result also shows a significant difference on user's perception of the information shown in the original and improved Sequence Logo where the user will comprehend the logo as no nucleotide present at position 1 on the original Sequence Logo design. The mean and standard deviation for the original Sequence Logo is $M=4.09$, $SD=1.405$ and for the improved Sequence Logo is $M=2.16$, $SD=0.660$ where the $t(54)=9.443$. The significant value for the performance is $p=0.000$ which is less than 0.05 ($p < 0.05$). Therefore, H_0 is rejected. This shows that the blank space at position 1 of the original Sequence Logo will cause the user to misinterpret as no nucleotide present in the logo compared to the information shown on the improved Sequence Logo.

Nevertheless, the results show that there is no significant difference in the amount of information shown in the original and improved Sequence Logo. The mean and standard deviation for the original Sequence Logo is $M=3.71$, $SD=1.227$ and for the improved Sequence Logo is $M=3.95$, $SD=1.113$ where the $t(54)=-1.066$. The significant value for the performance is $p=0.291$ which is greater than 0.05 ($p > 0.05$). Therefore, H_0 is accepted. This result shows that user will perceive both original and improved Sequence Logo, do not provide an adequate amount of information in the graphical representation.

4.4.2.4 Overall Preference

In addition, the paired sample t-test is also used to analyse user's preference on the use of the logo to represent the results. The null hypothesis that can be generated is:

H₀: There is no significant difference in the user's preference for using the original or improved Sequence Logo to represent the results.

Table 4.7 shows the result of the paired sample t-test between the two graphical representations.

Table 4.7: Paired sample t-test on the user's preference for graphical representation.

No.	Preference of graphical representation	Sequence Logo	Mean	Std. Deviation	t	df	Sig. (2-tailed)
1	As the representation of the results	Original design	5.78	0.567	11.084	54	0.000
		Improved design	4.15	0.870			
2	To differentiate the nucleotide A,T,G,C	Original design	4.49	0.742	1.479	54	0.145
		Improved design	4.18	1.203			
3	To identify the sequence conservation (from highest to lowest)	Original design	5.24	0.637	1.137	54	0.261
		Improved design	5.07	0.690			
4	As the graphical representation for paper publication	Original design	3.61	0.913	3.188	54	0.002
		Improved design	3.02	0.913			

The results above show that there is a significant difference in the preference of using the original or improved Sequence Logo as the representation of the result. The mean and standard deviation for the original Sequence Logo is $M=5.78$, $SD=0.567$ and for the improved Sequence Logo is $M=4.15$, $SD=0.870$ where the $t(54)=11.084$. The significant value for the performance is $p=0.000$ which is less than 0.05 ($p < 0.05$). Therefore, H_0 is rejected. The result shows that the user still prefers the original design of the Sequence Logo as the graphical representation for the results. This can be seen when then means value for the original Sequence Logo ($M=5.78$) is higher than the improved Sequence Logo ($M=4.15$) design.

Besides that, the results also show a significant difference on the preference of using the original or improved Sequence Logo design as the graphical representation of the result

in paper publication. The mean and standard deviation for the original Sequence Logo is $M=3.62$, $SD=0.912$ and for the improved Sequence Logo is $M=3.02$, $SD=0.912$ where the $t(54)=3.188$. The significant value for the performance is $p=0.002$ which is less than 0.05 ($p < 0.05$). Therefore, H_0 is rejected. The result shows that user still prefers the original design of the Sequence Logo as the representation of the result in paper publication compare to the improved Sequence Logo design.

Nonetheless, the result also shows no significant difference on the preference of using colour and alphabet shape in the original design or the arrangement of the nucleotide in rows on the improved Sequence Logo design to differentiate the nucleotide A, T, G, C in the graphical representation. The mean and standard deviation for the original Sequence Logo is $M=4.49$, $SD=0.742$ and for the improved Sequence Logo is $M=4.18$, $SD=0.1.203$ where the $t(54)=1.479$. The significant value for the performance is $p=0.145$ which is greater than 0.05 ($p > 0.05$). Therefore, H_0 is accepted. This shows that there is no difference on the preference of the user to use the colour and alphabet or the arrangement of the nucleotide in rows to differentiate each nucleotide in the graphical representation.

Other than that, the result also shows no significant difference on the preference of using the size and arrangement in the original design or the colour gradient on the improved Sequence Logo design to identify the sequence conservation in the graphical representation. The mean and standard deviation for the original Sequence Logo is $M=5.24$, $SD=0.637$ and for the improved Sequence Logo is $M=5.07$, $SD=0.690$ where the $t(54)=1.137$. The significant value for the performance is $p=0.261$ which is greater than 0.05 ($p > 0.05$). Therefore, H_0 is accepted. The result shows that there is no difference in the preference of the user to use the size and arrangement or the colour gradient to identify the sequence conservation in the graphical representation.

4.4.3 Evaluation on the perception of the improved Sequence Logo design based on user's skill (Objective 5)

In this evaluation, an independent sample t-test is used to analyse the perception of the user in assessing the context of the improved graphical representation based on different user's skill in using Sequence Logo. Table 4.8 shows the level of skills based on the Dreyfus novice expert scale (Dreyfus, Dreyfus, & Zadeh, 1987).

Table 4.8: Level of skills based on Dreyfus novice expert scale

Skills level	Stage	How context is assessed	Decision Making
1-2	Novice		
3-4	Advanced Beginner	Analytically	Rational
5-6	Competent		
7-8	Proficient	Holistically	Intuitive
9-10	Expert		

Based on the results from the previous evaluation study, it was found that there is a difference in how a user with different skill level assess or visualize the graphical representation of the original Sequence Logo design. Hence this evaluation will test on the difference in how a user with a different level of skill perceive the improved Sequence Logo design in terms of colour of the nucleotide, arrangement of the nucleotide, the information in the graphical representation and the overall design of the improved Sequence Logo.

An independent t-test is used to find the difference of how the graphical representation and information present in the improved Sequence Logo is being assessed between the analytic and holistic group users.

Table 4.9: Frequency & Percentage of analytic & holistic group user

Stage	Frequency	Percentage (%)	How context is assessed	Percentage (%)
Novice	6	10.91	Analytically	45.5
Advanced	9	16.36		
Beginner				
Competent	10	18.18	Holistically	54.5
Proficient	18	32.73		
Expert	12	21.82		
TOTAL	55	100	TOTAL	100

4.4.3.1 Colour of the nucleotide

The following shows the results of how analytic and holistic group user assess the colour of the nucleotide in the improved Sequence Logo.

The null hypothesis that can be generated is:

H₀: There is no significant difference on how the context of colour is being assessed in the improved Sequence Logo between analytically (skills level 1-6) and holistically (skills level 7-10) group.

Table 4.10 shows the result of the independent t-test between the two skill groups.

Table 4.10: Independent t-test of how the colour of the nucleotide in the improved Sequence Logo is being assessed.

Colour in the improved Sequence Logo design	Skills level	Mean	Std. Deviation	t	df	Sig. (2-tailed)
Identify the level of conservation based on the colour of the nucleotide bar	1-6 (analytically)	5.12	0.637	-0.192	53	0.849
	7-10 (holistically)	5.17	0.689			

Table 4.10: continued

The use of blue colour is suitable to represent the conservation values	1-6 (analytically)	4.58	4.44	-1.282	53	0.206
	7-10 (holistically)	5.50	4.10			
Able to differentiate the colour difference for the less conserved nucleotide	1-6 (analytically)	5.58	2.84	0.302	53	0.764
	7-10 (holistically)	3.31	2.77			

The results above show that there is no significant difference on how the context of colour is being assessed in the improved Sequence Logo between analytically and holistically group to identify the level of conservation based on the colour of the nucleotide bar. The mean and standard deviation for the analytic group is $M=5.12$, $SD=0.637$ and for the holistic group is $M=5.17$, $SD=0.689$ where the $t(53)=-0.192$. The significant value for the performance is $p=0.849$ which is greater than 0.05 ($p > 0.05$). Therefore, H_0 is accepted. This result shows that both analytic and holistic group are able to assess the level of conservation based on the four transition of blue colour in the improved design.

Besides that, the result also shows no significant difference on how the context of blue colour is used to represent the conservation value of the nucleotide in the improved Sequence Logo between analytically and holistically group. The mean and standard deviation for the analytic group is $M=4.44$, $SD=0.917$ and for the holistic group is $M=4.10$, $SD=1.029$ where the $t(53)=1.282$. The significant value for the performance is $p=0.206$ which is greater than 0.05 ($p > 0.05$). Therefore, H_0 is accepted.

Moreover, the result shows no significant difference on how the analytic and holistic group user can distinguish the difference in colour for the less conserved nucleotide. The mean and standard deviation for the analytic group is $M=2.84$, $SD=0.850$ and for the holistic group is $M=2.77$, $SD=0.935$ where the $t(53)=0.302$. The significant value for the

performance is $p=0.764$ which is greater than 0.05 ($p > 0.05$). Therefore, H_0 is accepted. The result shows that both group users disagrees that they are able to distinguish the colour difference for the less conserved nucleotide in the improved Sequence Logo.

4.4.3.2 Arrangement of the nucleotide

Besides the colour, the arrangement of the nucleotide in the improved design is also being evaluated. The following shows the results of how analytic and holistic group user assess the arrangement of the nucleotide in the improved Sequence Logo.

The null hypothesis that can be generated is:

H₀: There is no significant difference on how the arrangement of the nucleotide is being assessed in the improved Sequence Logo between analytically (skills level 1-6) and holistically (skills level 7-10) group.

Table 4.11 shows the result of the independent t-test between the two skill groups.

Table 4.11: Independent t-test of how the arrangement of the nucleotide in the improved Sequence Logo is being assessed.

The arrangement of a nucleotide in the improved Sequence Logo design	Skills level	Mean	Std. Deviation	t	df	Sig. (2-tailed)
Tend to view the nucleotide horizontally instead of vertically	1-6 (analytically)	5.20	0.707	0.520	53	0.605
	7-10 (holistically)	5.10	0.712			
Spacing between the nucleotide bar should be wider	1-6 (analytically)	4.56	0.768	-0.799	53	0.428
	7-10 (holistically)	4.73	0.828			
The dotted line can guide the user to view the nucleotide vertically	1-6 (analytically)	3.08	0.909	0.760	53	0.451
	7-10 (holistically)	2.90	0.845			

The results show that there is no significant difference on how the analytic and holistic group user visualize the arrangement of the nucleotide by viewing horizontally instead of vertically. The mean and standard deviation for the analytic group is $M=5.20$, $SD=0.707$ and for the holistic group is $M=5.10$, $SD=0.712$ where the $t(53)=0.520$. The significant value for the performance is $p=0.605$ which is greater than 0.05 ($p > 0.05$). Therefore, H_0 is accepted. The result shows that both group users agrees that they tend to visualize the logo horizontally (left to right) instead of vertically (top to bottom) with the arrangement of the nucleotide in rows on the improved design.

Besides that, the result also shows no significant difference on how the analytic and holistic group user assesses the spacing between the nucleotide bar to separate the nucleotide at each position in the improved design. The mean and standard deviation for the analytic group is $M=4.56$, $SD=0.768$ and for the holistic group is $M=4.73$, $SD=0.828$ where the $t(53)=-0.799$. The significant value for the performance is $p=0.428$ which is greater than 0.05 ($p > 0.05$). Therefore, H_0 is accepted.

Additionally, the result also shows no significant difference on how the use of dotted line on the left column of the logo can guide the user's eye movement to view from top to bottom instead of left to right between the analytic and holistic group user. The mean and standard deviation for the analytic group is $M=3.08$, $SD=0.909$ and for the holistic group is $M=2.90$, $SD=0.845$ where the $t(53)=0.760$. The significant value for the performance is $p=0.451$ which is greater than 0.05 ($p > 0.05$). Therefore, H_0 is accepted. The result shows that both group users disagrees that use of dotted line on the left column can help to guide their eye movement to view the logo vertically instead of horizontally.

4.4.3.3 Information on the improved Sequence Logo

In addition, the context of information shown in the improved Sequence Logo design is also being evaluated. The following shows the results of how analytic and holistic group user assess the information shown in the improved Sequence Logo.

The null hypothesis that can be generated is:

H₀: There is no significant difference on how the information display in the graphical representation is being assessed between analytically (skills level 1-6) and holistically (skills level 7-10) group.

Table 4.12 shows the result of the independent t-test between the two skill groups.

Table 4.12: Independent t-test of how the information is being assessed in the improved Sequence Logo.

Assess of information in the improved Sequence Logo	Skills level	Mean	Std. Deviation	t	df	Sig. (2-tailed)
The '<' indication helps the user to comprehend the conservation value as less than 0.01 bit	1-6 (analytically)	4.48	0.963	-1.505	53	0.138
	7-10 (holistically)	4.87	0.937			
The detailed bits value helps to easily identify the information content	1-6 (analytically)	4.52	4.146	4.146	53	0.000
	7-10 (holistically)	3.23	4.138			
The total number of aligned sequence helpful for analysis of results	1-6 (analytically)	5.68	0.802	0.220	53	0.826
	7-10 (holistically)	5.63	0.765			
The consensus sequence helps to easily identify the most conserved nucleotide at a position	1-6 (analytically)	5.16	0.898	3.173	53	0.003
	7-10 (holistically)	4.20	1.270			

Table 4.12: continued

Too many information present on the representation	1-6 (analytically)	4.88	0.781	-0.968	53	0.337
	7-10 (holistically)	5.10	0.884			
Get more information about the relationship of the nucleotide at each position	1-6 (analytically)	4.28	1.595	0.684	53	0.497
	7-10 (holistically)	4.00	1.438			

The results show that there is no significant difference on how the analytic and holistic group user assess the information shown at position 1 of the improved design where the ‘<’ symbol is used to indicated the conservation value of the nucleotide as less than 0.01 bits. The mean and standard deviation for the analytic group is $M=4.48$, $SD=0.963$ and for the holistic group is $M=4.86$, $SD=0.937$ where the $t(53)=-1.505$. The significant value for the performance is $p=0.138$ which is greater than 0.05 ($p> 0.05$). Therefore, H_0 is accepted. The result shows that both group users agrees that they use of the ‘<’ symbol helps them to easily comprehend the value of conservation as less than 0.01 bits.

Besides that, the result shows a significant difference on how the analytic and holistic group user assesses the bits value shown on top of the nucleotide bar to indicate the information content of each nucleotide in the improved Sequence Logo. The mean and standard deviation for the analytic group is $M=4.52$, $SD=1.159$ and for the holistic group is $M=3.23$, $SD=1.135$ where the $t(53)=4.146$. The significant value for the performance is $p=0.000$ which is less than 0.05 ($p< 0.05$). Therefore, H_0 is rejected. The result shows that the analytic group agree that the detailed bits value show on top of the nucleotide bar can help them to easily identify the information content of the nucleotide easily.

Other than that, the result shows no significant difference on how the information about the total number of aligned sequence in the improved design can the analytic and

holistic group user analyse the result. The mean and standard deviation for the analytic group is $M=5.68$, $SD=0.802$ and for the holistic group is $M=5.63$, $SD=0.765$ where the $t(53)=0.220$. The significant value for the performance is $p=0.826$ which is greater than 0.05 ($p > 0.05$). Therefore, H_0 is accepted. The result shows that both group users agrees that the information about the number of the aligned sequences used to generate the Sequence Logo is useful for them to analyse the results.

Nevertheless, the result shows a significant difference on how the analytic and holistic group user assesses the consensus sequence written at the top of the graphical representation to identify the most conserved nucleotide at each position. The mean and standard deviation for the analytic group is $M=5.16$, $SD=0.898$ and for the holistic group is $M=4.20$, $SD=1.270$ where the $t(53)=3.173$. The significant value for the performance is $p=0.003$ which is less than 0.05 ($p < 0.05$). Therefore, H_0 is rejected. This result shows that the analytic group agree that the consensus sequence written on top of the improved graphical representation can help them to easily identify the most conserved nucleotide.

Furthermore, the result shows no significant difference on how the analytic and holistic group user assesses the amount of information shown on the improved Sequence Logo. The mean and standard deviation for the analytic group is $M=4.88$, $SD=0.781$ and for the holistic group is $M=5.10$, $SD=0.884$ where the $t(53)=-0.968$. The significant value for the performance is $p=0.337$ which is greater than 0.05 ($p > 0.05$). Therefore, H_0 is accepted. The result shows that both group users agrees that there is too many information shown on the improved graphical representation which causes the logo to look too complicated compared to the original Sequence Logo.

Moreover, the result also shows no significant difference on how the analytic and holistic group user assesses the information about the relationship between the nucleotide at each position. The mean and standard deviation for the analytic group is $M=4.28$, $SD=1.595$ and for the holistic group is $M=4.00$, $SD=1.438$ where the $t(53)=0.684$. The significant value for the performance is $p=0.497$ which is greater than 0.05 ($p > 0.05$). Therefore, H_0 is accepted. The result shows that both group users agrees that the arrangement of the nucleotide in rows can help them to visualize the relationship of the same nucleotide at a different position.

4.4.3.4 Overall Design

Meanwhile, the perception of both analytic and holistic group on the overall design of the improved Sequence Logo is also evaluated. The null hypothesis that can be generated is:

H₀: There is no significant difference on the way analytically (skills level 1-6) and holistically (skills level 7-10) group use the improved Sequence Logo as an extra reference to interpret the results but remain the use of Original Sequence Logo as the main graphical representation of the result.

Table 4.13 shows the result of the independent t-test between the two skill groups.

Table 4.13: Independent t-test of how analytic and holistic group user will use the improved Sequence Logo to represent the result.

The use of Improved Sequence Logo	Skills level	Mean	Std. Deviation	t	df	Sig. (2-tailed)
Use the improved Sequence Logo as an extra reference for me to interpret the results but remain the use of Original Sequence Logo as the main graphical representation	1-6 (analytically)	5.12	0.781	-6.498	53	0.000
	7-10 (holistically)	6.27	0.521			

The result shows a significant difference on how the analytic and holistic group user will use the improved Sequence Logo as an extra reference for me to interpret the results but remain the use of Original Sequence Logo as the main graphical representation. The mean and standard deviation for the analytic group is $M=5.12$, $SD=0.781$ and for the holistic group is $M=6.27$, $SD=0.521$ where the $t(53)=-6.498$. The significant value for the performance is $p=0.000$ which is less than 0.05 ($p < 0.05$). Therefore, H_0 is rejected. This result shows that the holistic group user agrees that they will remain the use of the original Sequence Logo design to interpret the result whereas the improved design can be used as an extra reference for analysis.

4.5 Discussion

The results from the evaluation above show that although the improved Sequence Logo design can provide more information for the user, the original design of the Sequence Logo remains to be the preference for the user to represent the results due to some limitations on the design of the improved Sequence Logo.

In terms of the use of colour, the use of the blue colour transition in the improved Sequence Logo can help the user to identify the level of conservation, however, problems on the choice of colour used for the less conserved nucleotide is found. In the improved

graphical representation, light blue colour is used to represent the most conserved nucleotide at a position whereas dark blue colour is used to represent the least conserved nucleotide. The finding of the result shows that problems are found in the representation for less conserved nucleotide where user is unable to differentiate the transition of the darker colour if the height of the nucleotide is too short. Therefore, more attention is needed for the user to identify the non-conserved nucleotide sequence compare to the original design of Sequence Logo. Hence, an improvement on the choice of transition colour used to represent the non-conserved nucleotide is needed in order to reduce the amount of attention to identifying the sequence conservation on the improved design.

Besides that, the finding from the evaluation also shows that enhancement of the arrangement for the improved graphical representation is needed to assist the user in interpreting the context of the improved logo. Limitation on the way user visualizes the improved Sequence Logo are found where the user tends to view the logo from left to right (horizontally) rather than from top to bottom (vertically) which causes more attention needed to identify the level of conservation at a position. Based on the visual hierarchy model, searching and scanning of the context are carried out during the visualization process (Soussan, Siegel, & Tullis, 2011). The searching refers to the finding of the entry point whereas the scanning is the extraction of information around the entry point (Soussan, Siegel, & Tullis, 2011). The visual hierarchy model suggest that during both processes, the visual attributes in the context such as the size, colour, contrast, alignment, repetition, proximity, white space, texture and style, and etc. are important in influencing the way user visualize the context (Soussan, Siegel, & Tullis, 2011). If these attributes are used effectively, it will create a visual hierarchy that will guide the user's eye movement while viewing the page. However, due to the use of dotted line on the left column are too far from the nucleotide

stack across the logo and the spacing between the nucleotide at each position across the graphical representation is smaller than the whitespace between the rows of nucleotide, user will tend to perceive the improved Sequence Logo in 'E' manner (Figure 10). Hence, more attention is needed to visualize the level of conservation at a position compared to the original design of the logo.

Other than that, the results show that the arrangement of the nucleotide in rows in the improved Sequence Logo design helps the user to easily identify the height of each nucleotide at each position easily compare to the original Sequence Logo. However, improvement is still needed as the improved design also causes the loss of information about the total height of a stack of the nucleotide at a position to compare to the original design.

Moreover, an improvement on the amount of information shown in the improved Sequence Logo is needed where most of the user found that there is too many information shows on the improved Sequence Logo which causes the graphical representation to be more complicated than the original design. Besides that, enhancement on the design of representation for the similar conservation nucleotide at a position is needed to help the user in interpreting the conservation of nucleotide if the nucleotide has the same relative frequency.

Besides that, the result also shows that although the improved Sequence Logo is able to provide more information and better for visualizing the conservation value, the original design is still preferred by the user for representation of the results due to the complicated design of the improved Sequence Logo. Therefore, improvement to simplify the design is needed to help the user in visualizing all the information in the graphical representation.

4.6 Summary

The results from this evaluation study has shown that the improved Sequence Logo can provide more information to the users to interpret the result however there are still some limitations on the improved design in terms of the choices of colour for the nucleotides, the arrangement of the nucleotide where the gap between each column is too large that causes the loss of information about the total height and the amount of information shown is found to be too much. Therefore future enhancement is needed to improve the design again to solve these problems. The next chapter will discuss furthermore on future works that can be done to improve the Sequence Logo design and the limitation found during the study.

CHAPTER 5

CONCLUSION & FUTURE WORKS

5.1 Overview

This chapter will summarize the finding of the study and discusses the limitation found during the study. Besides that, suggestion for the future works that can be done to improve the design of the Sequence logo is also discussed in this chapter.

5.2 Summary of the Findings

5.2.1 Evaluation on the visual attributes performance on the design of the Sequence Logo based on the human factors principles in perception and visualization (Objective 1)

The result from the study shows that the limitation on the colour of the nucleotide used was standardization of the colour to represent each nucleotide is needed to avoid confusion to the user. Standard Sequence Logo generator allows the user to change the colour of the nucleotide-based on their preference (Lee & Oon, 2013). This problem will cause confusion especially for the novice user who does not have experience in using the logo to misinterpret the information especially if the shape of the nucleotide are unable to identify if the size is too small. Colour is often used to identify the non-conserved nucleotide in the logo when the size of the symbol is too hard to visualize. Therefore, improvement to apply a standardize colour on the nucleotide is needed to avoid further confusion.

Besides that, the design of the logo must keep as simple as possible without any line indication for the bits value. Instead of a line, pop up indication is a better choice to enable users to visualize the bits value of the nucleotide. Hence, interactivity can be added to the

logo generator to enhance the visualization where the logo can be kept in a simple design at the same time display the detailed information when needed.

Sequence Logo is frequently used as a graphical representation for comparison of different multiple sequence alignment. The comparison is frequently done by glancing the logo back and forth which requires high cognitive load from the user and misinterpretation will occur if the height is very similar to each other (Sakai & Aerts, 2014). Therefore improvement on the design where the user is allowed to compare the nucleotide at each position easily is needed to allow users to easily compare a few logos together.

5.2.2 Evaluation on the difference between novice and experts while interpreting the result from Sequence Logo based on the Dreyfus novice expert skill acquisition model (Objective 2)

The result from the study shows that there is a difference on the way novice and expert user interpret the result where novice user who has less experience prefer more detailed information to assist them in analysing the result whereas expert user does not need too detailed information to help them make a decision. The different way of perception and making the decision are affected by the level of experience where novice user who does not frequently use the logo will assess the representation in analytically whereas expert user tends to perceive the representation in a holistic manner.

Hence, improvement to display the amount of information that is suitable for both level user is needed so that novice user will not misinterpret the results. This can be done by adding interactive design such as the pop-up bits value in the tool to hide and show the detailed information based on the user needs.

5.2.3 Evaluation based on user's perception of the design of the original and improved graphical representation (Objective 4)

The result from the study shows that the choice of colour used to represent the conserved and non-conserved nucleotide is very important in the graphical representation. Few factors such as the size of the nucleotide, the light source, the type of display such as on printed paper or monitor display should be considered as a factor that will cause the different perception of the graphical representation. These factors must be taken as a consideration while choosing the colour especially when it is visualized on the monitor is very different from the colour printed out on paper. Therefore improvement in the choice of colour that can be easily distinguished while these factors exist is needed to avoid misperception by the user.

Besides that, the arrangement of the nucleotide at each position is also very important in the design of Sequence Logo. The different user needs a different type of information for their analysis. Some user would like to view the height of each nucleotide whereas some just want to view the total height. It is hard to accommodate the needs of the different user in a graphical representation. Therefore, interactivity on the tools can be added to enable the user to choose the information they wanted to display on the representation.

5.2.4 Evaluation on the difference of perception between novice and experts while interpreting the improved Sequence Logo design (Objective 5)

The results of the study show that improvement on the colour, arrangement and information representation on the Sequence Logo is needed to assist the user with a different level of skill and experience to analyse the graphical representation without any biases or misinterpretation. However, it is important to choose the correct visual attribute and the amount of information to represent the context in the graphical representation. The results

from the second evaluation study show that the more information shown on a graphical representation, the more attention is needed for the user to interpret the results.

Although the colour used in the improved Sequence Logo can help the user to identify the level of conservation for the nucleotide. However, an improvement on the choice of colour is needed as there are other factors to be considered while choosing the colour for the representation such as the height of the nucleotide bar and the size of the graphical representation. Other than that, the factor on the visual hierarchy is also needed to be considered during the design of the Sequence Logo to improve the arrangement of the nucleotide.

An expert user who is used to the original design might not easily get used to the new Sequence Logo design compare to the novice user. Therefore time of adjustment is needed and further improvement that can suit the needs of the expert user is needed to enable all users to benefit from the improvement on the Sequence Logo. One of the solutions for this problem is to add interactivity on the tools to generate Sequence Logo where the different user can choose different representation that is suitable for their research.

5.3 Limitation of the Study

Some limitation was found during the data collection in this study where the response rate for the online survey is very low despite more than 200 participants were being invited to participate in the survey. This problem occurs due to the population of the user that has experience in using Sequence Logo are found through the paper publication or university website where the email in the publication or website may not be updated or exist anymore if the student graduated from the university. Besides that, the time for data collection is

longer for overseas participants as they can only be contacted through emails and academician or researcher may respond that they are too busy to respond to the survey.

Additionally, besides the improvement based on the visual attributes issues found from the first evaluation study, some factors should also be considered during the design of the improved Sequence Logo such as the size of the logo, the choice of colour displayed on different medium (paper or monitor) and the algorithm used to calculate consensus sequence. These factors must be taken into account for the reason user still prefer the original design of the graphical representation. Hence, further improvement is needed to solve these problems to design a better graphical representation.

5.4 Future Works

The results of the finding show that there are a few aspects to improve such as:

- The enhancement of colour to represent the conservation level is needed in order to help the user to easily distinguish the different level of conservation in a different medium for the improved Sequence Logo.
- The improvement on the arrangement of the nucleotide is needed where the attributes that will affect the visual hierarchy must be considered while designing the logo in order to aid user's eye movement to visualize the important information in the graphical representation.
- The improvement in the amount of information shown on the improved graphical representation is needed. Although the improved design can display more information about the conservation value to the user it was found that the graphical representation will become more complicated than the original Sequence Logo. Therefore, interactivity on the tools to generate the Sequence Logo can be added for

the user to choose the type of information to be shown in the graphical representation. This will help both novice and expert user to select the most suitable graphical representation needed for different purposes such as the analysis of result or as a graphical representation of the result in the paper publication.

5.5 Contribution of Data Visualization

The exponentially increasing amount of data in the field of biology, financial, climate, health care, World Wide Web and etc. especially with the emerging of Big Data has increased the importance of data visualization tools in the market. How to effectively represent the huge amount of data in a single graphical representation has become an essential problem for visual designer to design a graphical representation that can efficiently display a large amount of information to people to run their business, solve their problem and assist in decision making (Ward, Grinstein, & Keim, 2015). Biologist uses visualization tools to display the sequence alignment to find a specific sequence for transcription factor binding site, stock analyst visualize the stock market chart to analyse the market trends, climate researcher visualize the changes of rainfall volume to predict flood area, medical researcher visualize the spread of disease in different area, all these usages of graphical representation shows how important it is for data visualization to correctly convey the message to user as the wrong display of data might directly affect the life of a person (Sravanthi & Reddy, 2015).

Problems such as visual noise, information loss, high rate of data change, a high performance requirement, perceptual and interactive scalability are the challenges visual designer need to face every day while designing a graphical representation (Wang, Wang, & Alexander, 2015). Hence, a continuous study on data visualization is very essential for a

visual scientist to provide effective solutions for these problems in order to create a graphical representation that can efficiently convey the messages to the user.

REFERENCES

- Alexandre, D. S., & Tavares, J. M. (2010). Introduction of Human Perception in Visualization. *International Journal of Imaging and Robotics*, 4, 60-70.
- Bailey, T. L. (2008). Discovering Sequence Motifs. In J. M. Keith (Ed.), *Bioinformatics* (pp. 231-251). New Jersey: Humana Press.
- Brandenburg, E., Zander, S., Figge, A., & Beier, G. (2014, July). Recommendations for Tracelink Decisions—An Empirical Investigation of Visualizations. In Ben Amaba und Brian Dalgetty (Eds.): *Advances in Human Factors, Software and Systems Engineering. 5th International Conference on Applied Human Factors and Ergonomics: AHFE Conference* (pp. 3-11).
- Chauhan, R., & Agarwal, P. (2012). A Review: Applying Genetic Algorithms for Motif Discovery. *International Journal of Computer Technology & Applications*, 3(4), 1510-1515.
- Creswell, J. W. (2013). *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage publications.
- Crooks, G. E., Hon, G., Chandonia, J.-M., & Brenner, S. E. (2004). Weblogo: A Sequence Logo Generator. *Genome Research*, 14(6), 1188-1190.
- Difference Between Cognition and Perception. (2014, December 23). Retrieved from DifferenceBetween.com:<http://www.differencebetween.com/difference-between-cognition-and-vs-perception/>
- Dreyfus, H. L., Dreyfus, S. E., & Zadeh, L. A. (1987). Mind over Machine: the power of human intuition and expertise in the age of the computer. *IEEE Expert*, 2(2), 110-111.
- Dreyfus, S. E., & Dreyfus, H. L. (1980). A five-stage model of the mental activities involved in directed skill acquisition. California University Berkeley Operations Research Center.

- Few, S. (2004, September 18). *Enie, meenie, minie, moe: selecting the right graph for your message*. Retrieved from Perceptual Edge:
https://www.perceptualedge.com/articles/ie/the_right_graph.pdf
- Few, S. (2004, September 4). *Tapping the power of visual perception*. Retrieved from Perceptual Edge: https://www.perceptualedge.com/articles/ie/visual_perception.pdf
- Few, S. (2013). Data Visualization for Human Perception. In *The Encyclopedia of Human-Computer Interaction* (2nd ed.). Retrieved from Interaction Design Foundation.
- Friendly, M. (2008). A Brief History of Data Visualization. *Handbook of Computational Statistics: Data Visualization*. Springer-Verlag, 15-56.
- Goldstein, E. B. (2007). *Sensation & Perception* (7th ed.). Canada: Vicki Knight.
- Han, S. H., Humphreys, G. W., & Chen, L. (1999). Uniform connectedness and classical Gestalt principles of perceptual grouping. *Perception & Psychophysics*, 61(4), 661-674.
- Hastie, R., & Dawes, R. M. (2010). *Rational choice in an uncertain world: The psychology of judgment and decision making* (2nd ed.). California: Sage Publication.
- Healey, C. G., & Enns, J. (2012). Attention and Visual Memory in Visualization and Computer Graphics. *IEEE Transactions on Visualization & Computer Graphics*, 18(7), 1170-1188.
- Huang, L. Q., & Palsher, H. (2007). A Boolean map theory of visual attention. *Psychological Review*, 114(3), 599.
- Khatri, N., & Ng, H. A. (2000). The role of intuition in strategic decision making. *Human Relation*, 53(1), 57-86.
- Koofka, K. (2013). *Principles of Gestalt Psychology* (Vol. 44). London: Routledge.

- Kultys, M., Nicholas, L., Schwarz, R., Goldman, N., & King, J. (2014). Sequence Bundles: A novel method for visualising, discovering and exploring sequence motifs. *3rd IEEE Symposium on Biological Data Visualization*. 8. BioMed Central Proceedings.
- Lee, N. K., & Oon, Y. B. (2012). Decision Making Biases in Using Sequence Logo Visualization. *Southeast Asian Network of Ergonomics Societies Conference (SEANES)*. Langkawi: IEEE.
- Lee, N. K., & Oon, Y. B. (2013). Potential Perils of Biological Sequence Visualization using Sequence Logo. *Computer Graphics, Imaging and Visualization (CGIV) 10th International Conference*. IEEE.
- Li, W., Yang, B., Liang, S., Wang, Y., Whiteley, C., Cao, Y., & Wang, X. (2008). BLogo: A Tool For Visualization of Bias in Biological Sequences. *Bioinformatics*, 24, 2254-2255.
- Lin, I. (2012). *Discovering Transcription Factor Binding Motif Sequence*. Stanford University, Department of Biology. Bioc218 Final Report.
- Maguire, E., Philippe, R.-S., Sansone, S.-A., & Chen, M. (2014). Redesigning the Sequence Logo with Glyph-based. *Eurographics Conference on Visualization*.
- Nisbett, R. E., & Yuri, M. (2005). The influence of culture: holistic versus analytic perception. *Trends in Cognitive Sciences*, 9(10), 467-473.
- Nisbett, R. E., Peng, K., Incheol, C., & Ara, N. (2001). Culture and Systems of Thought: Holistic vs. Analytic Cognition. *Psychological Review*, 108, 291-310.
- Norman, H. (2007). *How Experts Differ from Novice*. Retrieved from The Source book for Teaching Science: <https://www.csun.edu/science/ref/reasoning/how-students-learn/2.html>
- O'Donoghue, S. I., Gavin, A.-C., Gehlenborg, N., Goodsell, D. S., Hériché, J.-K., Nielsen, C. B., North, C., Olson, A. J., Procter, J. B., Shattuck, D. W., Walter, T., & Wong, B. (2010). Visualizing biological data-now and in the future. *Nature Method*, S2-S4.

- Olsen, L. R., Kudahl, U. J., Simon, C., Sun, J., Schönbach, C., Reinherz, E. L., Zhang, G. L., & Brusic, V. (2013). BlockLogo: Visualization of peptide and sequence motif conservation. *Journal of Immunological Methods*, 400, 37-44.
- Padmanabhan, U., & Choudhury, S. (2013, July 15). *How to Make Data Visualization Better with Gestalt Laws*. Retrieved from Six Revisions: <http://sixrevisions.com/usability/data-visualization-gestalt-laws/>
- Petre, M., & Green, T. R. (1993). Learning to read graphics: some evidence that 'seeing' an information display is an acquired skill. *Journal of Visual Languages & Computing*, 4(1), 55-70.
- Phillips, J. K., Gary, K., & Sieck, W. R. (2004). Expertise in judgment and decision making: A case for training intuitive decision skills. In *Blackwell handbook of judgment and decision making* (pp. 297-315).
- Randel, J. M., Pugh, L. H., & Reed, S. K. (1996). Differences in expert and novice situation awareness in naturalistic decision making. *International Journal of Human-Computer Studies*, 45(5), 579-597.
- Ray, W. C., Rumpf, R. W., Sullivan, B., Callahan, N., Magliery, T., Machiraju, R. W., Wong, B., Krzywinski, M. & Bartlett, C. W. (2014). Understanding the sequence requirements of protein families: Insights from the BioVis 2013 contests. *Biomed Central*.
- Sakai, R., & Aerts, J. (2014). Sequence Diversity Diagram for comparative. *3rd IEEE Symposium on Biological Data Visualization*. 8. BioMed Central Proceedings.
- Schneider, T. D., & Stephens, M. R. (1990). Sequence Logos: a new way to display consensus sequences. *Nucleic Acids Research*, 18(20), 6097-6100.

- Djamasbi, S., Siegel, M., & Tullis, T. (2011). Visual hierarchy and viewing behavior: An eye tracking study. *Human-Computer Interaction. Design and Development Approaches*, 6761, 331-340.
- Stone, M. (2006, January 17). *Choosing colours for data visualization*. Retrieved from Perceptual Edge: https://www.perceptualedge.com/articles/b-eye/choosing_colors.pdf
- Vacic, V., Lakoucheva, L. M., & Radivojac, P. (2006). Two Sample Logo: A Gaphical Representation of the Differences Between Two Sets of Sequence Alignments. *Bioinformatics*, 22, 1536-1537.
- Weirauch, M. T., Cote, A., Norel, R., Annala, M., Zhao, Y., Riley, T. R., Julio, S., Cokelaer, T., Vedenko, A., Talukder, S., Bussemaker, H. J., Morris, Q. D., Hughes, T. R., & Stolovitzky, G. (2013). Evaluation of methods for modeling transcription factor sequence specificity. *Nature Biotechnology*, 31(2), 126-134.
- Wong, B. (2010). Gestalt Principles (Part 1). *Nature Methods*, 7(11), 863-941.
- Zhu, Y. (2007). Measuring Effective Data Visualization. *Advances in Visual Computing*, 4842, 652-661.

APPENDICES

Appendix 1

Evaluation Study on the Visualization of Biological Sequence Motif

Page 1

This questionnaire consists of three sections, Section A, Section B, and Section C. Please complete all the questions in each section to reflect your opinion(s) on current visualization of Sequence Logo for DNA & RNA sequences.

Section A: Demographics

Please **tick (✓)** on the box given.

1. Gender

☐ Male

☐ Female

2. Age

_____ years old

3. Occupation

☐ Researcher

☐ Academician

☐ Other _____

☐ Graduate Student

☐ Undergraduate Student

4. Areas of expertise

5. Please rate your skills in using Sequence Logo.

Novice				Competent				Expert	
1	2	3	4	5	6	7	8	9	10

6. What are the tools that you used to visualize DNA or RNA sequence motifs? (tick all that apply)

☐ Weblogo

☐ MATLAB

☐ Other _____

☐ RNA Structure Logo

☐ None

☐ GENIO/logo

Page 2
Section B

The following questions are based on the visualization of sequence logo generated by Weblogo. Please complete all the questions below. Figure 1.0 shows an example of Sequence Logo for NHLH1 from JASPAR database.



Figure 1.0 Example of the visualization of a sequence logo.

a. Colour

Strongly Disagree	Disagree	Slightly Disagree	Neutral	Slightly Agree	Agree	Strongly Agree
1	2	3	4	5	6	7

1.	I can more easily identify the nucleotide with the default colours (green, red, orange and blue) respectively compared to alternative black and white colours for A, T/U, G, and C.	1	2	3	4	5	6	7
2.	There is no need for user to customize the colour of the nucleotides	1	2	3	4	5	6	7
3.	The use of different colours help user to identify the non-conserved bases that exist in the sequence.	1	2	3	4	5	6	7
4.	It is necessary to standardize the colours of the nucleotides.	1	2	3	4	5	6	7
5.	It is easier to identify the consensus sequence if the colour of the most dominant nucleotide at each position are the same. (see Figure 2.0)	1	2	3	4	5	6	7

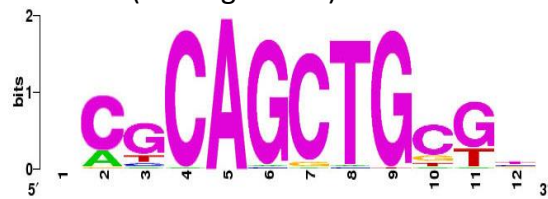
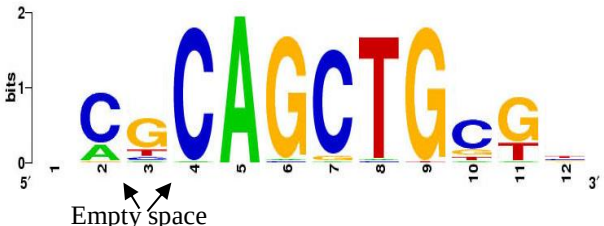
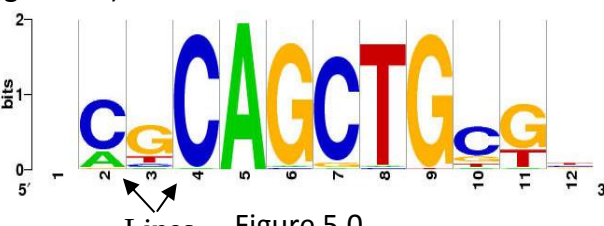


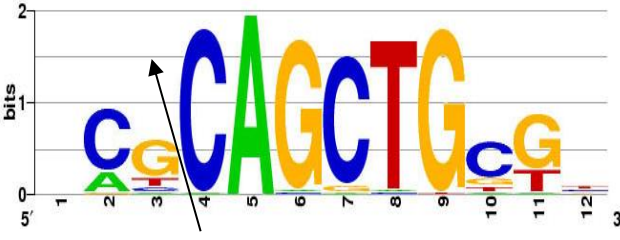
Figure 2.0

b. Layout

Strongly Disagree	Disagree	Slightly Disagree	Neutral	Slightly Agree	Agree	Strongly Agree
1	2	3	4	5	6	7

1.	The arrangement of the consensus sequence on the top most position helps user to easily identify the conserved nucleotides.	1	2	3	4	5	6	7
2.	It is easier to identify the conserved nucleotides if the characters for the conserved nucleotides are made bold. (see Figure 3.0)	1	2	3	4	5	6	7
 Figure 3.0								
3.	The arrangement of the nucleotides at each position is too near to one another (without empty space between each position) which result in a messy looking sequence logo.	1	2	3	4	5	6	7
4.	I prefer empty spaces to exist in the immediate adjacent positions of every stack of nucleotide in the sequence logo. (see Figure 4.0)	1	2	3	4	5	6	7
 Figure 4.0								
5.	Putting a thin line to separate every stack of nucleotide adjacent to each other will result in a neater and more organized sequence logo. (see Figure 5.0)	1	2	3	4	5	6	7
 Figure 5.0								

6.	The vertical lines on the background, indicating the vertical axis bit values, helps user to determine the relative conservation value (in bits), of each nucleotide. (see Figure 6.0)	1	2	3	4	5	6	7
----	--	---	---	---	---	---	---	---



Vertical Lines

Figure 6.0

Page 4

c. Decision making

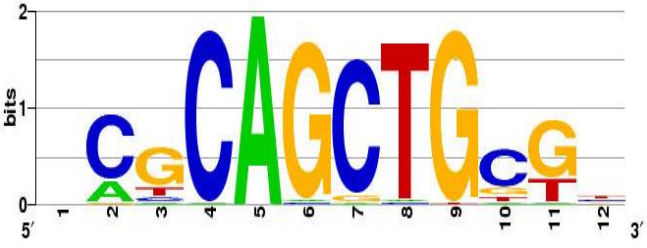
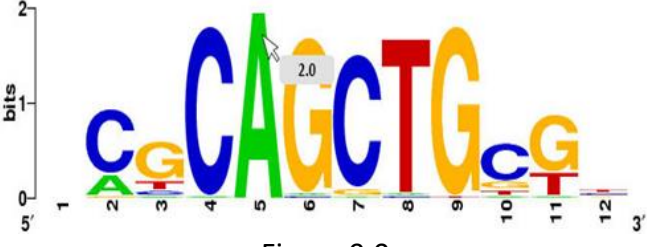
Strongly Disagree	Disagree	Slightly Disagree	Neutral	Slightly Agree	Agree	Strongly Agree
1	2	3	4	5	6	7

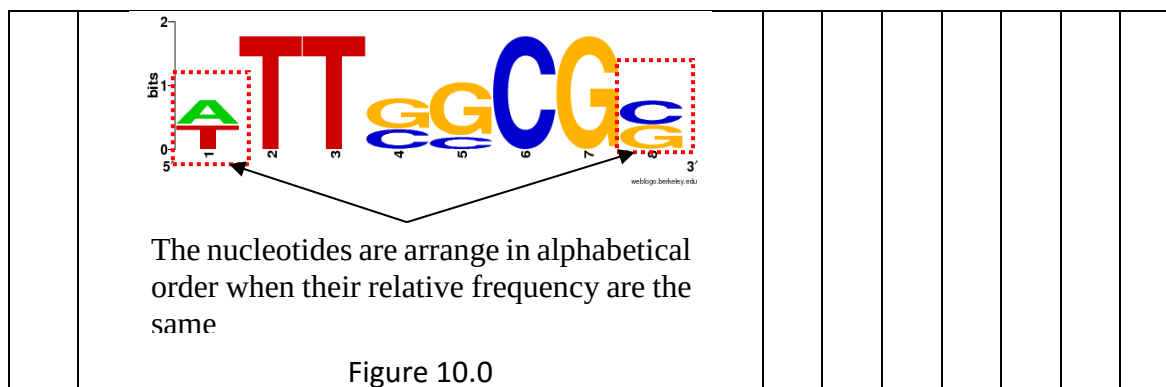
1.	The information shown on a sequence logo is perfectly sufficient for a user to identify and compare the consensus sequence.	1	2	3	4	5	6	7
2.	I can determine the dominant level of the nucleotides based on the available information.	1	2	3	4	5	6	7
3.	I do not need the value of the relative frequency of the nucleotides on the sequence logo.	1	2	3	4	5	6	7
4.	The 'bits' value on the y-axis is useful to me for identifying the conserved nucleotide.	1	2	3	4	5	6	7
5.	There is no need to show the nucleotides if their relative frequency are the same at the same position, e.g. at position 1 in Figure 7.0.	1	2	3	4	5	6	7



Figure 7.0

6.	It is hard to decide which nucleotide is more conserved if their relative frequencies are almost identical.	1	2	3	4	5	6	7
----	---	---	---	---	---	---	---	---

7.	The vertical lines at the background of the logo helps user to quickly decide which of the nucleotides are more important. (see Figure 8.0)  Figure 8.0	1	2	3	4	5	6	7
8.	I prefer the bits value of a nucleotide character to be shown to the user when the cursor is hovered over it. (see Figure 9.0)  Figure 9.0	1	2	3	4	5	6	7
9.	I can easily interpret the information content of nucleic acids in a sequence logo by verifying their height.	1	2	3	4	5	6	7
10.	I believe that the sequence logo generated by Weblogo is very accurate.	1	2	3	4	5	6	7
11.	I have never misinterpreted the information presented in sequence logo generated by the tools.	1	2	3	4	5	6	7
12.	I will normally compare a few sequence logos generated by several tools in order to get the most accurate result.	1	2	3	4	5	6	7
13.	I prefer to use the information content-based logo (sequence logo) compared to the frequency logo even though the frequency logo can more clearly shows the height of each nucleotides present.	1	2	3	4	5	6	7
14.	I will normally interpret the sequence logo with many conserved nucleotides as more superior compare to the sequence logo with only a few conserved nucleotides.	1	2	3	4	5	6	7
15.	It will be misleading to the user when the nucleotides are arranged by alphabetical order when both of their relative frequencies are identical. (For example the A & G nucleotides at position 1 and C & T at position 11 in Figure 10.0)	1	2	3	4	5	6	7



d. Overall satisfaction

Strongly Disagree	Disagree	Slightly Disagree	Neutral	Slightly Agree	Agree	Strongly Agree
1	2	3	4	5	6	7

1.	I think the sequence logo is a very good graphical representation for visualizing sequence motif.	1	2	3	4	5	6	7
----	---	---	---	---	---	---	---	---

Section C

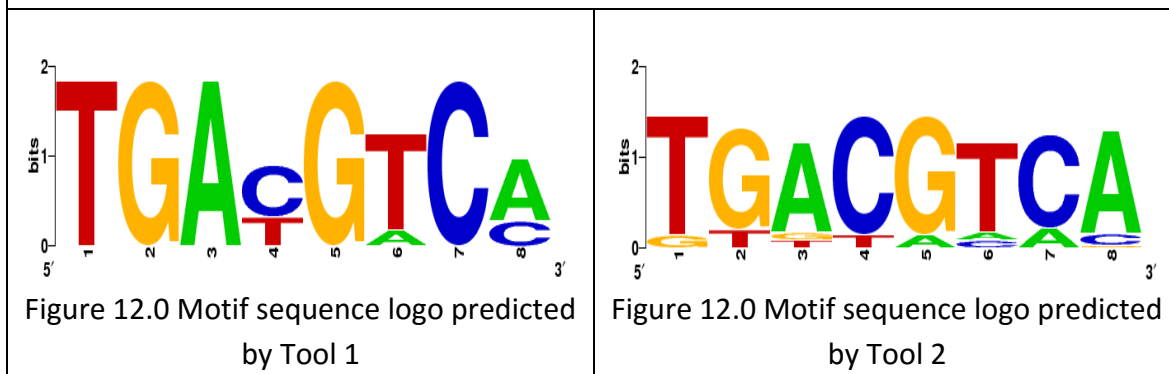
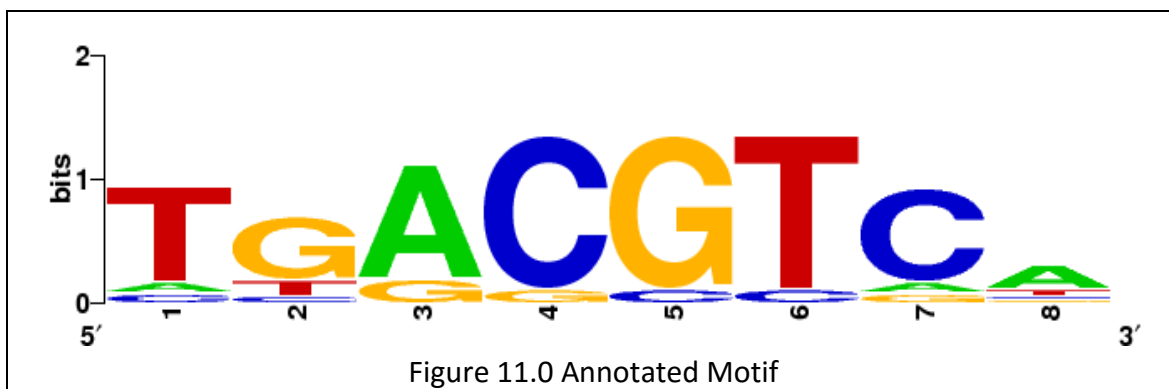
Please complete **all** the questions below.

- i) Please rate the importance of each of the following piece of information provided in a sequence logo.

Very Unimportant	Unimportant	Slightly Unimportant	Neutral	Slightly Important	Important	Very Important
1	2	3	4	5	6	7

1.	Consensus Sequence	1	2	3	4	5
2.	Non-conserved bases	1	2	3	4	5
3.	Position of the aligned sequences	1	2	3	4	5
4.	Error bars	1	2	3	4	5
5.	Nucleotide's colour	1	2	3	4	5
6.	Value of bits	1	2	3	4	5
7.	Relative frequency of each nucleotide	1	2	3	4	5

- ii) Figures below show the annotated motif (Figure 11.0) and the motifs predicted computationally by Tool 1 (Figure 12.0) and Tool 2 (Figure 13.0).



Based on your perception, compare the sequence motif in Figure 12.0 and Figure 13.0 with the annotated motif Figure 11.0 and rate the **accuracy** of both predicted logos.

Very Inaccurate				Very Accurate			
1	2	3	4	5	6	7	

Very Inaccurate				Very Accurate			
1	2	3	4	5	6	7	

Based on your perception, compare the sequence motif in Figure 12.0 and Figure 13.0 with the annotated motif (Figure 11.0) and rate the **importance** of the nucleotide **G** at position 3 in both predicted logos.

Negligible				Very Important			
1	2	3	4	5	6	7	

Negligible				Very Important			
1	2	3	4	5	6	7	

Based on your perception, compare the sequence motifs in Figure 12.0 and Figure 13.0 with the annotated motif Figure 11.0 and rate the **quality** of the predicted motif in both logos.

Very Low				Very High			
1	2	3	4	5	6	7	

Very Low				Very High			
1	2	3	4	5	6	7	

Open Ended Questions

1. Please provide an example on how you would usually use sequence logos and/or for what purposes you use the sequence logos.
2. Do you face any problem(s) or limitation(s) in interpreting the graphical representation when you use sequence logo in your research? If you answer YES, please provide a scenario or an example of the problem(s)?
3. What other information(s) do you need from the sequence logo for your research?
4. Is there any other recommendation(s) that you would like to suggest to improve on the visualization of sequence logos?

Appendix 2

Evaluation Study on the Visualization of Improved Sequence Logo Design

Page 1

This questionnaire consists of 2 sections, Section A, and Section B. Please complete all the questions in each section to reflect your opinion(s) on the current visualization of Sequence Logo and the improved design of Sequence Logo.

Section A: Demographics

Please **tick (✓)** on the box given.

1. Gender

☐

Male

☐

Female

2. Age

_____ years old

3. Occupation

☐

Researcher

☐

Academician

☐

Student : Undergrad / Postgrad

☐

Others, please list _____

4. Areas of expertise

5. Please rate your skills in using Sequence Logo.

Novice			Competent				Expert		
1	2	3	4	5	6	7	8	9	10

6. What are the tools that you used to visualize DNA or RNA sequence motifs? (tick all that apply)

☐

Weblogo

☐

RNA Structure Logo

☐

GENIO/logo

☐

Seq2Logo

☐

None

☐

Other _____

7. I usually use Sequence Logo:

☐

As a graphical representation for my results in paper publication

☐

To do comparison between a few Sequence Logos to visualize the conservation characteristics of the multiple sequence alignment

☐	To get the conservation value of the nucleotide found at each position in a Sequence Logo
☐	Others: _____
☐	All of the above

Page 2

Section B: Comparison between the current design and the improved Sequence Logo design

The following shows the comparison between the original design of the Sequence Logo (Figure 1.0) and the improved design of Sequence Logo (Figure 2.0) for NHLH1 sequence alignment. Please answer all the questions below.

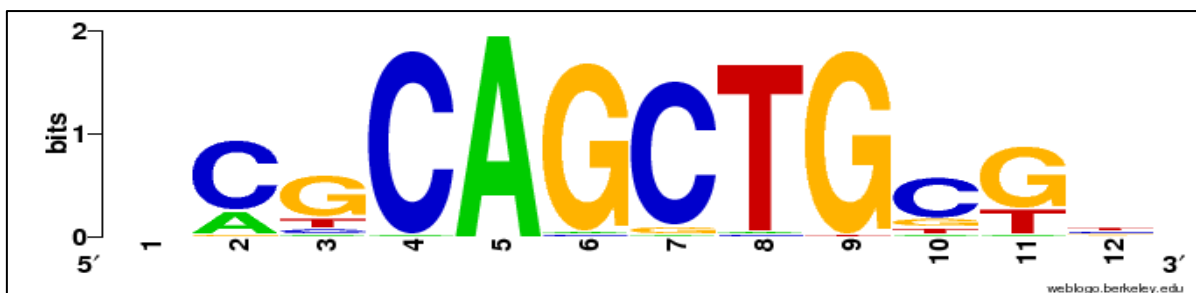


Figure 1.0 Original design of Sequence Logo for NHLH1

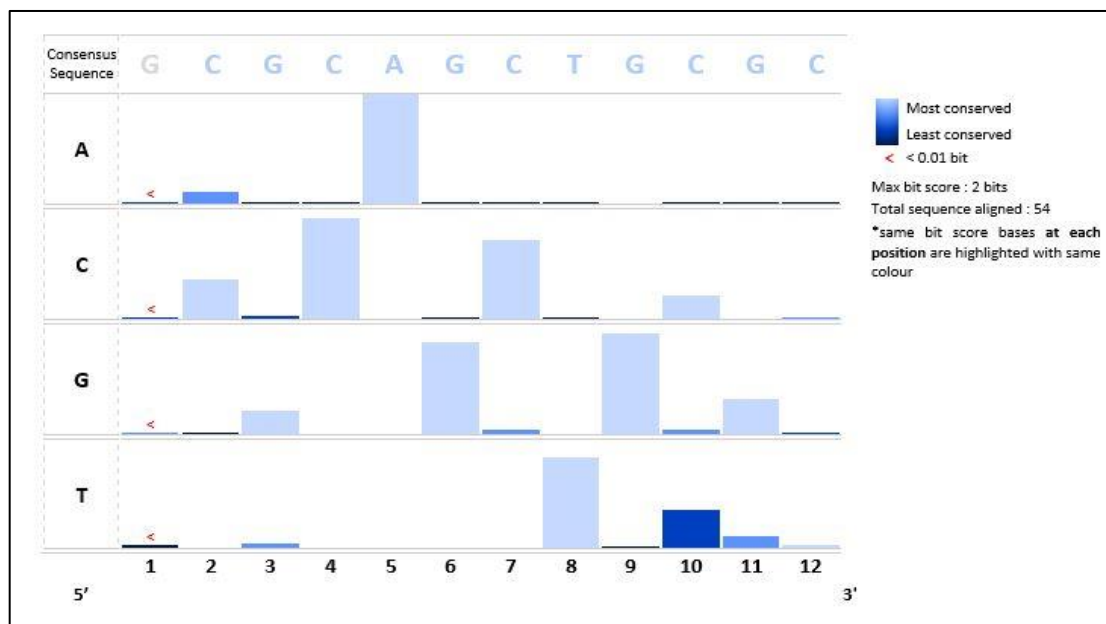


Figure 2.0 Improved design of Sequence Logo for NHLH1

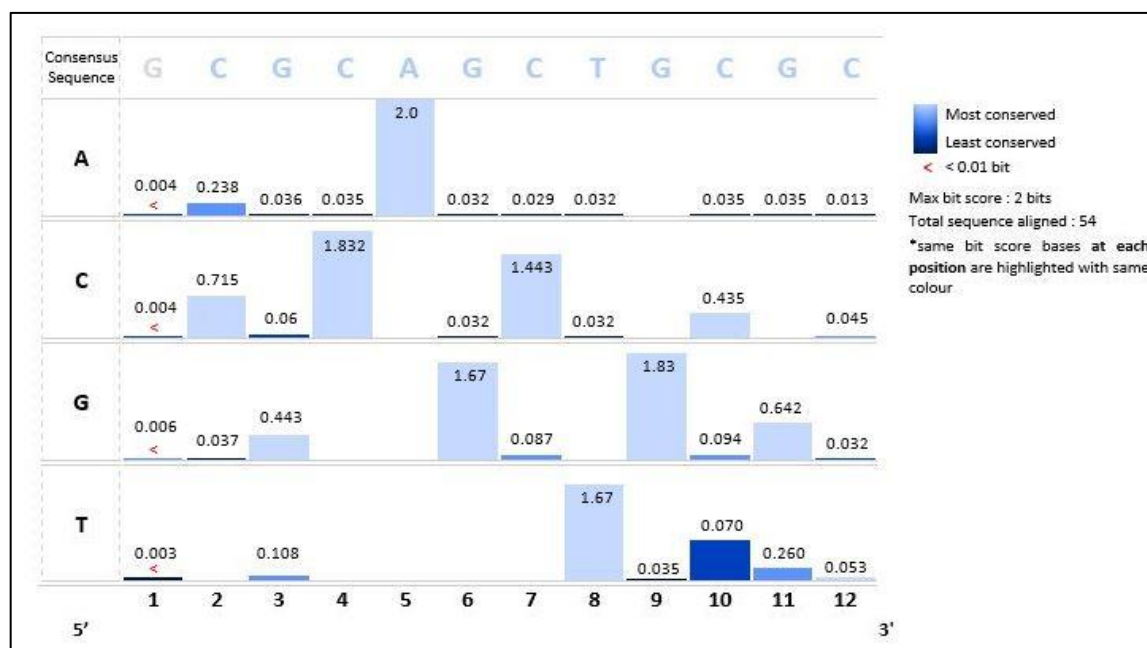


Figure 3.0 Improved design of Sequence Logo with detailed bits value

a. Nucleotide's Colour

Strongly Disagree	Disagree	Slightly Disagree	Neutral	Slightly Agree	Agree	Strongly Agree
1	2	3	4	5	6	7

1. I use the colour of the nucleotide in the Original Sequence Logo to:

a.	Differentiate the nucleotide A,T,G,C	1	2	3	4	5	6	7
b.	Identify the consensus sequence	1	2	3	4	5	6	7
c.	Identify the non-conserved sequence	1	2	3	4	5	6	7
d.	Identify the level of conservation of the nucleotide.	1	2	3	4	5	6	7

2. I use the colour of the nucleotide in the Improved Sequence Logo to:

a.	Differentiate the nucleotide A,T,G,C	1	2	3	4	5	6	7
b.	Identify the consensus sequence	1	2	3	4	5	6	7
c.	Identify the non-conserved sequence	1	2	3	4	5	6	7
d.	Identify the level of conservation of the nucleotide.	1	2	3	4	5	6	7

3. I can easily distinguish the level of conservation in the Improved Sequence Logo based on the colour of the nucleotide bars.

1	2	3	4	5	6	7
---	---	---	---	---	---	---

4. The use of blue colour is suitable to represent the conservation values in the Improved Sequence Logo design.

1	2	3	4	5	6	7
---	---	---	---	---	---	---

5. I am able to differentiate the colour difference for the less conserved nucleotide in the Improved Sequence logo.

1	2	3	4	5	6	7
---	---	---	---	---	---	---

6. Colour Preference

a.	I prefer the representation of colour in the Original Sequence Logo to differentiate the nucleotide.	1	2	3	4	5	6	7
b.	I prefer the representation of colour in the Improved Sequence Logo to identify the conservation level of the nucleotide.	1	2	3	4	5	6	7

Page 3

- b. Nucleotide's Arrangement

Strongly Disagree	Disagree	Slightly Disagree	Neutral	Slightly Agree	Agree	Strongly Agree
1	2	3	4	5	6	7

1. The arrangement of the nucleotide in a stack in the Original Sequence Logo **does not** helps me to identify the **height of each nucleotide** present.

1	2	3	4	5	6	7
---	---	---	---	---	---	---

2. The arrangement of the nucleotide in the improved Sequence logo helps me to identify the **height of each nucleotide** easily.

1	2	3	4	5	6	7
---	---	---	---	---	---	---

3. The arrangement of the nucleotide shown in helps me to easily identify the **total height** of the nucleotide stack at each position

a.	Original Sequence Logo	1	2	3	4	5	6	7
b.	Improved Sequence Logo	1	2	3	4	5	6	7

4. I can **compare the height** of a nucleotide present at each position in easily.
E.g. the height of nucleotide **A** present at each position.

a.	Original Sequence Logo	1	2	3	4	5	6	7
b.	Improved Sequence Logo	1	2	3	4	5	6	7

5. The stacking arrangement of the nucleotide at each position shown in helps me to identify which nucleotide **present and not present** at a position easily.

a.	Original Sequence Logo	1	2	3	4	5	6	7
b.	Improved Sequence Logo	1	2	3	4	5	6	7

6. I can easily comprehend the arrangement of the nucleotide at **position 1** in as all the nucleotide are present with similar conservation value.

a.	Original Sequence Logo	1	2	3	4	5	6	7
b.	Improved Sequence Logo	1	2	3	4	5	6	7

7. The arrangement of the nucleotide A, C, G, T help me to easily identify the consensus sequence.

a.	Original Sequence Logo	1	2	3	4	5	6	7
b.	Improved Sequence Logo	1	2	3	4	5	6	7

8. I tend to view the nucleotide from left to right (horizontally) than top to bottom (vertically) in the Improved Sequence Logo

1	2	3	4	5	6	7
---	---	---	---	---	---	---

9. The spacing between the nucleotide at each position should be wider to separate the nucleotide at each position in the Improved Sequence Logo.

1	2	3	4	5	6	7
---	---	---	---	---	---	---

10. The dotted line on the left column of the Improved Sequence Logo guide me to view the nucleotide from top to bottom (vertically).

1	2	3	4	5	6	7
---	---	---	---	---	---	---

Page 4

c. Information provided

Strongly Disagree	Disagree	Slightly Disagree	Neutral	Slightly Agree	Agree	Strongly Agree
1	2	3	4	5	6	7

1. The information provided in are very adequate for me to identify the conservation value and make decision based on the graphical representation.

a.	Original Sequence Logo	1	2	3	4	5	6	7
b.	Improved Sequence Logo	1	2	3	4	5	6	7

2. I can easily identify **how many sequence alignments were used** to generate the logo by visualizing the

a.	Original Sequence Logo	1	2	3	4	5	6	7
b.	Improved Sequence Logo	1	2	3	4	5	6	7

3. It is **not necessary** to visualize the nucleotide present at **position 1 in the Original Sequence Logo** since their conservation value are very **low**.

1	2	3	4	5	6	7
---	---	---	---	---	---	---

4. It will cause confusion to user if the nucleotide present at **position 1** are not shown in the Original Sequence Logo.

1	2	3	4	5	6	7
---	---	---	---	---	---	---

5. The '**<**' indication in the Improved Sequence Logo at position 1 helps me to easily comprehend the conservation value as less than 0.01 bit.

1	2	3	4	5	6	7
---	---	---	---	---	---	---

6. I will comprehend the has **no nucleotide** present at **position 1** of the alignment.

a.	Original Sequence Logo	1	2	3	4	5	6	7
b.	Improved Sequence Logo	1	2	3	4	5	6	7

7. The detailed bits value shown in the Improved Sequence Logo helps me to identify the information content of the nucleotide easily.

1	2	3	4	5	6	7
---	---	---	---	---	---	---

8. The information about the total number of aligned sequence on the right of the Improved Sequence Logo design is useful for me to analyse the results.

1	2	3	4	5	6	7
---	---	---	---	---	---	---

9. I can understand the information shown in the Improved Sequence Logo easily by reading the legend on the right of the graphical representation

1	2	3	4	5	6	7
---	---	---	---	---	---	---

10. The information about the consensus sequence at the top of the Improved Sequence Logo helps me to easily identify the consensus sequence from the graphical representation.

1	2	3	4	5	6	7
---	---	---	---	---	---	---

11. I do not need the information about the bits value to be shown on the graphical representation in the Improved Sequence Logo to identify the conservation level of the nucleotide.

1	2	3	4	5	6	7
---	---	---	---	---	---	---

12. There are too many information shown on the Improved Sequence Logo which makes the graphical representation looks too complicated compare with the Original Sequence Logo.

1	2	3	4	5	6	7
---	---	---	---	---	---	---

13. I can get more information about the relationship between the nucleotide at each position in the Improved Sequence Logo. E.g. The relationship between nucleotide A at each position.

1	2	3	4	5	6	7
---	---	---	---	---	---	---

Page 5

d. Overall visualization

Strongly Disagree	Disagree	Slightly Disagree	Neutral	Slightly Agree	Agree	Strongly Agree
1	2	3	4	5	6	7

1. I prefer the use of as the graphical representation of my results.

a.	Original Sequence Logo	1	2	3	4	5	6	7
b.	Improved Sequence Logo	1	2	3	4	5	6	7

2. I prefer the use of to differentiate the nucleotide A, C, G & T.

a.	Colour & Alphabet Shape in the Original Sequence Logo	1	2	3	4	5	6	7
b.	Arrangement in rows from top to bottom in the Improved Sequence Logo	1	2	3	4	5	6	7

3. The use of helps me to easily identify the **sequence conservation (highest to lowest)** in the graphical representation

a.	Size & Arrangement (most conserved to least conserved) in the Original Sequence Logo	1	2	3	4	5	6	7
b.	Colour gradient (light blue to dark blue) in the Improved Sequence Logo	1	2	3	4	5	6	7

4. I will use the Improved Sequence Logo as an extra reference for me to interpret the results but remain the use of Original Sequence Logo as the main graphical representation.

1	2	3	4	5	6	7
---	---	---	---	---	---	---

5. The is a suitable graphical representation in paper publication to represent the results of the sequence alignment.

a.	Original Sequence Logo	1	2	3	4	5	6	7
b.	Improved Sequence Logo	1	2	3	4	5	6	7

Appendix 3

List of publication & conference proceeding:

1. Kok, W. Y., Oon, Y. B., & Lee, N. K. (2016). Perception enhancement using visual attributes in sequence motif visualization. *BioRxiv*.doi:10.1101/066928
2. Kok, W.Y., Oon, Y.B., Lee, N.K. (2014). Sequence Logo Visualization based on Gestalt Perception–Novices vs Experts. 3rd International Conference of the South East Asian Network of Ergonomics Societies. Singapore: IEEE.

List of conference paper presentation:

1. Kok, W. Y., Oon, Y. B., & Lee, N. K. Perception enhancement using visual attributes in sequence motif visualization. In 3rd International Conference on Interactive Digital Media, 2-4 December 2014, Sutera Harbour, Sabah, Malaysia.
2. Kok, W., Oon, Y. B., & Lee, N. K. Sequence Logo Visualization based on Gestalt Perception–Novices vs Experts. In 3rd International Conference of the South East Asian Network of Ergonomics Societies, 1-4 December 2014, Holiday Inn Singapore Atrium, Singapore.