2012 Southeast Asian Network of Ergonomics Societies Conference (SEANES)

Decision Making Biases in using Sequence Logo Visualization

Nung Kion, Lee[1], Yin Bee, Oon[2]
Faculty of Cognitive Sciences and Human Development
Universiti Malaysia Sarawak
Kota Samarahan, 94300 Sarawak
{[1]nklee, [2]yinbee}@fcs.unimas.my

***Abstract*— Sequence Logo is a visualization method for displaying conservations characteristics of a sequence (DNA, RNA, proteins) motif profile obtained from either wet-lab or computational analysis. Usage of visualization in decision making carries some elements of subjectivity. In addition, people's decisions are often biased in favor of their proposed hypotheses. The objectives of this paper were to examine the biases in using sequence logo as an evaluation metric for transcription factor analysis and identify some critical weaknesses in sequence logo for possible future improvements. Document analysis and subject matter expert interviews method were used for information gathering. We found that sequence logo has been frequently misused to support the results obtained from computational transcription factor analysis. In addition, we suggest that current sequence logo can be improved in several aspects to support various users' needs and minimize elements of subjectivity in decision making.**

I. INTRODUCTION

In scientific studies, visualization is important for communicating methodology or empirical result to readers. It aims to improve the quality and clarity of information presentation. Sequence logo [1] has been a popular and widely accepted method for visualizing motif composition of biological sequences in the past 20 years. A motif is the characteristic feature of transcription factor (TF) protein binding sites or short segments in protein/RNA sequences that have specific functional roles in cells. In this paper, we focus our study on DNA motif. A probabilistic motif profile of a TF, which is constructed from a set of real or putative binding sites, represents the specificity of binding sites it bound. The sequence logo shows the conservation property and relative frequencies of nucleotide symbols (i.e., A, C, G, T) in each position on a multiple-aligned binding sequences. This can facilitate the identification of a motif's characteristic signature. A sequence logo has typically used for several purposes in leading journal articles: (a) to illustrate motif characteristics obtained from wet-lab or computational analysis; (b) to compare and contrast motifs obtained in an evaluation study; or (c) as performance evaluation metric in various experimental studies including wet-lab or computational tools.

Despite being useful, the sequence logo has some apparent weaknesses. Since a logo only visualizes the summarized information in a motif, it has some associated risks when used

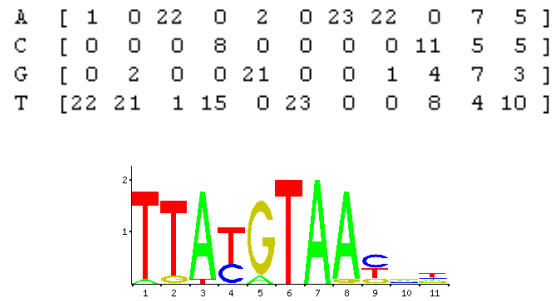

Figure 1. An example of a motif sequence logo. Shown is the NFIL3 motif profile obtained from JASPAR. The top sub-figure is the PFM of the multiple aligned sites whereas the bottom one is its sequence logo.

for objective comparison purposes. The interpretation of a logo might be subject to the researcher's confirmation biases [4]. For example, in computational motif prediction, the raw binding sites used to generate the sequence logo are not known to a reader. As such, performance comparison of computational motif prediction tools based on logo could lead to inaccurate claims. In addition, the use of sequence logos for comparison purposes is relied on individual perception and experiences and therefore the interpretation can be very subjective [6]. As a result, some conclusion deduced from a visualized motif logo can be misleading and portray inaccuracies when use in some contexts of studies. Other than that, there could be mismatch in terms of the amount of information convey to the reader and how the information is perceived. That is, visualizations may appear more convincing and sound than they really are [6]. For example, the use of colors in a sequence logo can impress the reader more than the actual quality of the motifs presented. These weaknesses can, to a certain extent, jeopardize the quality and precision needed in scientific result publications.

Several novel and improved versions, based on the original sequence logo visualization methods, have been proposed. Reference [7] proposed CorreLogo to visualize RNA/DNA logos in 3D. RNAlogo was proposed to visualize RNA motif [8]. Two Sample Logo has extended the original sequence logo to include visualizing of statistical differences between two sets of sequence alignment [9]. Reference [10] proposed BerryLogo, a motif visualization method based on the log-odd scores instead of the information content used in the sequence logo. enoLogo allows visualization of logo based on the

Both authors contributed equally in this study.

978-1-4673-1734-4/12/$31.00 ©2012 IEEE