2013 10th International Conference Computer Graphics, Imaging and Visualization

Potential Perils of Biological Sequence Visualization using Sequence Logo

Lee Nung Kion
Faculty of Cognitive Sciences & Human Development
Universiti Malaysia Sarawak
Kota Samarahan, Sarawak
Email: nklee@fcs.unimas.my

Oon Yin Bee
Faculty of Cognitive Sciences & Human Development
Kota Samarahan, Sarawak
Universiti Malaysia Sarawak
Email: yinbee@fcs.unimas.my

***Abstract*—Sequence motif's characteristics are commonly visualized by using a sequence logo. This paper describes a user study aimed at evaluating the effectiveness of sequence logo as evaluation metric for motif prediction tools. We also investigate the nature of confirmation biases in using sequence logos in result reporting in publications. While sequence logos have been widely used for visualizing sequence motifs in the past 20 years, no study has reported its effectiveness and possible misuses in decision making. We conducted a paper-and-pencil test to determine the effectiveness of sequence logos in some of their common usages. A survey study was also performed to investigate sequence logos' learnability. We found that there are great mismatches between users' perception and actual quality of motifs when sequence logos were used as an evaluation metric. Therefore, evaluation of motif prediction tools based on sequence logos has to be interpreted cautiously. Our result also suggests that there are still room for improvements in the current sequence logo's layout design.**

I. Introduction

Information visualization is a visual representation of raw or processed data in a more appealing way so that it improves communication to human in a more cognitive friendly manner. Due to the complexity of biological data, e.g., genes, 3D structural property, motifs, and often complex relationship between them, visualization is a very useful and powerful technique to present data in a more meaningful and comprehensible way. The ultimate aim is to amplify cognitive performance in various tasks. In the motif prediction problem, bioinformaticians are interested to discover over-represented motif patterns that are recurrent in a set of biological sequences (e.g, proteins or DNA). A sequence motif is a characteristic nucleotide or amino acid sequence that is conserved in a group of sequences. In most cases, it has a biological function. In this paper, we focus our investigation on the visualization of DNA motif. DNA motifs are functional elements located in the upstream or downstream of genes they regulate during the gene expression process. The interactions between transcription factor proteins and their binding sites, i.e. proteins-DNA interaction, determine the rate and when proteins are produced.

A sequence logo [1] graphically visualizes the intrinsic characteristics of motifs–the conservation of nucleotides, in DNA/RNA or proteins. A sequence logo of a sequence motif is constructed in three consecutive steps: (a) sequences are multiple aligned by using an ungapped multiple alignment tool such as CLUSTALW or MEME. Alignment columns are then trimmed on both ends to retain only the ungapped positions; (b) the aligned sequences are represented by a position-frequency-matrix (PFM) [2], which represents the likelihood of nucleotide $b \in \{A, C, G, T\}$ occurs at position i of a sequence motif, i.e., $p(b|i)$; (c) the sequence logo is generated from the PFM by using the information theory principle. Readers can refer to [3] for an illustrated example of how a sequence logo is generated.

The idea of a sequence logo was originated from the maximum information delivery of motif information for improving accuracy on motif analysis [4]. It was argued that, a sequence logo depicts more informative and accurate nucleotide compositions in aligned binding sites in comparison with a consensus string representation. Nevertheless, as far as we know, the human factor aspects of a sequence logo have not been considered in the design. Nor there is any study that investigates human factor issues relevant to a sequence logo.

A sequence logo highlights two critical pieces of nucleotides conservation information in a motif. The first is the conservation level in each multiple-alignment column of a motif which is measured in bits, with 2 bits as the maximum conservation for DNA sequences. The second is the relative frequency of the four nucleotides, i.e., A, C, G, T, which is represented by the total height of each symbol in a particular alignment position. In addition, from a sequence logo, we can identify possible minor or major grooves of the binding sequences of a TF.

The perils of visualization for communicating information have been reported in many works [5]. Studies have shown that the assessment on the quality of scientific studies seems to be particularly vulnerable to confirmation bias. In other words, scientists tend to rate studies that report findings consistent with their prior beliefs more favorably than studies reporting findings inconsistent with their previous beliefs.

In our previous study[6], we argued that there were severe confirmation biases when the sequence logo was used as computational tools evaluation metric. We found some claims reported in published articles were flawed because of the different heuristic rules employed for motif comparison. Furthermore, because visualized motif does not show the actual motif information, some hidden attributes about the quality of the

978-0-7695-5051-0/13 $26.00 © 2013 IEEE
DOI 10.1109/CGIV.2013.26