Sequence analysis

BLogo: A tool for visualization of bias in biological sequences

Wenchen Li¹, Bo Yang¹, Shaoguang Liang², Yonghua Wang³, Chris Whiteley⁴ Yicheng Cao¹ and Xiaoning Wang¹,⋆

¹School of Bioscience and Bioengineering, South China University of Technology, Guangzhu 510641, China,
²National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural University, Wuhan 430070, China,
³College of Light Industry and Food Sciences, Key Lab of Fermentation and enzyme engineering, South China University of Technology, Guangzhou 510641, China,
⁴Department of Biochemistry, Microbiology & Biotechnology, Rhodes University, Grahamstown 6139, South Africa

Associate Editor: Dr. Limsoon Wong

ABSTRACT
Summary: Blogo is a web-based tool that detects and displays statistically significant position-specific sequence bias with reduced background noise. The over-represented and under-represented symbols in a particular position are shown above and below the zero line. When the sequences are in open reading frames, the background frequency of nucleotides could be calculated separately for the three positions of a codon, thus greatly reducing the background noise. The z test or fisher’s exact test is used to evaluate the statistical significance of every symbol in every position and only those that are significant are highlighted in the resulting logo. The perl source code of the program is freely available and can be run locally.

Contact: lwcbio@yahoo.com.cn, xnwang@21cn.net

1 INTRODUCTION
Sequence Logo and WebLogo were created and developed, respectively, by Schneider and Stephens (1990) and Crooks et al., (2004) as ways to visualize sequence conservation. Though the sequence logo was used as a user-friendly generator there were two major drawbacks to note: (i) Every symbol was assumed to have equal distribution. When the sequences were from a biased genome (i.e. high G+C or A+T content), or open reading frames (ORFs) where the G+C content were different for the three positions of a codon, the sequence logo had a high background noise and the informative signal was not well defined (Hasan and Schreiber, 2006). (ii) The traditional sequence logos were designed to show “conservation” rather than “bias” of sequences. Sometimes one needed to study the over-represented or under-represented symbols in a region of sequences, which could not be presented in any traditional logo. We have developed a new sequence logo to overcome these two limitations, and by using it we have reported a study about the bias of nucleotides/amino acids in the 5'/N terminal of genes/proteins in prokaryotic genomes (Li et al., 2007). Here a website and a Perl source code have been created making the method freely available both online and for local installation.

2 METHODS
In a biased genome, the formula of information content was modified in at least two solutions. Schreiber (2002) and Hasan (2006) applied two concepts from IT, distortion and patterned interference (a type of noise) to correct signals. Gorodkin (1997) and Stormo (1998) calculated the information content relative to a background distribution. Here, the algorithm was in accord with that of Stormo’s.

Information content was calculated for each position of sequences using the formula:

\[ H_{(i)} = \sum H_{(i,L)} = \sum (P_{(i,L)} \log_2 \frac{P_{(i,L)}}{P_i}) \]

where \( L \) is the position in the sequences; \( i \) is symbols (A, T, C and G for nucleotides or 20 amino acids for protein sequences); \( P_{(i,L)} \) is the average probability of symbol \( i \) at position \( L \); and \( P_i \) is the background probability of symbol \( i \). \( H_{(i,L)} \) is positive when \( P_{(i,L)} \) is bigger than \( P_i \), and negative when \( P_{(i,L)} \) is smaller than \( P_i \).

For type 1 logo, the total height of the symbols in a position \( L \) is equal to \( H_{(i,L)} \) and the height of every symbol \( i \) is proportional to its observed frequency. For type 2 logo, all the letters from each stack are ordered from the biggest \( H_{(i,L)} \) to the smallest, and all letters with a positive \( H_{(i,L)} \) are stacked above zero. The height of a symbol \( i \) equals to \( H_{(i,L)} \).

For each symbol in every position, the \( \chi^2 \) test (in large samples) or fisher’s exact test (in small samples) was used to evaluate the statistical significance of the difference between the frequency of that symbol in that position and the background (expected). The difference was assumed to be significant if \( P \)-value was less than the threshold value (the default value was 0.05 and this can be modified by user).

3 INPUT AND GRAPHICAL OUTPUT

*To whom correspondence should be addressed.
Fig. 1. Logos created from the 21 nucleotides downstream of the start codon of 642 genes of *S. coelicolor* (an organism of high G+C content and unbalanced nucleotide contents for the three positions of codon). (a) Type 1 logo with setting the background frequency of A, G, C and T to 0.25. (b) Type 2 logo with the background frequency of nucleotides for the three positions of codon calculated separately.

The input sequences could be DNA, RNA, protein or codon in flat or fasta format. The background frequency of symbols could be calculated from the input sequences or by the user. When the sequences were codons, the background frequency of nucleotides could be calculated separately for the three positions of a codon. Blogo could create two types of logos: the type 1 logo was similar to WebLogo and the type 2 logo was designed to show sequence bias, for which the over-represented and under-represented symbols were stacked above and below the zero line, and the height of every symbol was proportional to its information content. The color of symbols was the same as the default setting of WebLogo, unless when the statistical test was used, a symbol with a P-value larger than a threshold (default is 0.05) was colored gray. In the case of a type 1 logo (Figure 1a), setting the equal background frequency of A, G, C and T made very high “noise”, especially for the third position of the codon in which the nucleotide frequency were highly unequal. On the contrary, an example of the type 2 logo (Figure 1b) showed that the nucleotides A, C and T are over-represented and C, T, G under-represented respectively for the three positions (numbered 1, 2 and 3) of the first codon. This information was not shown with the type 1 logo (Figure 1a).

**FUNDING**

China Postdoctoral Science Foundation (20070410239 to W.L.); National Natural Science Foundation of China (20506007, 20706021); the Research Fund for the Doctoral Program of Higher Education (20070561073).

**ACKNOWLEDGEMENTS**

We thank the Computing Cluster of South China University of Technology.

**REFERENCES**


