DNAlogo: a smart mini application for generating DNA sequence logos

Yabin Guo[1*]

Affiliation:

[1]Guangdong Provincial Key Laboratory of Malignant Tumor Epigenetics and Gene Regulation, Medical Research Center, Sun Yat-sen Memorial Hospital, Sun Yat-sen University, Guangzhou, China.

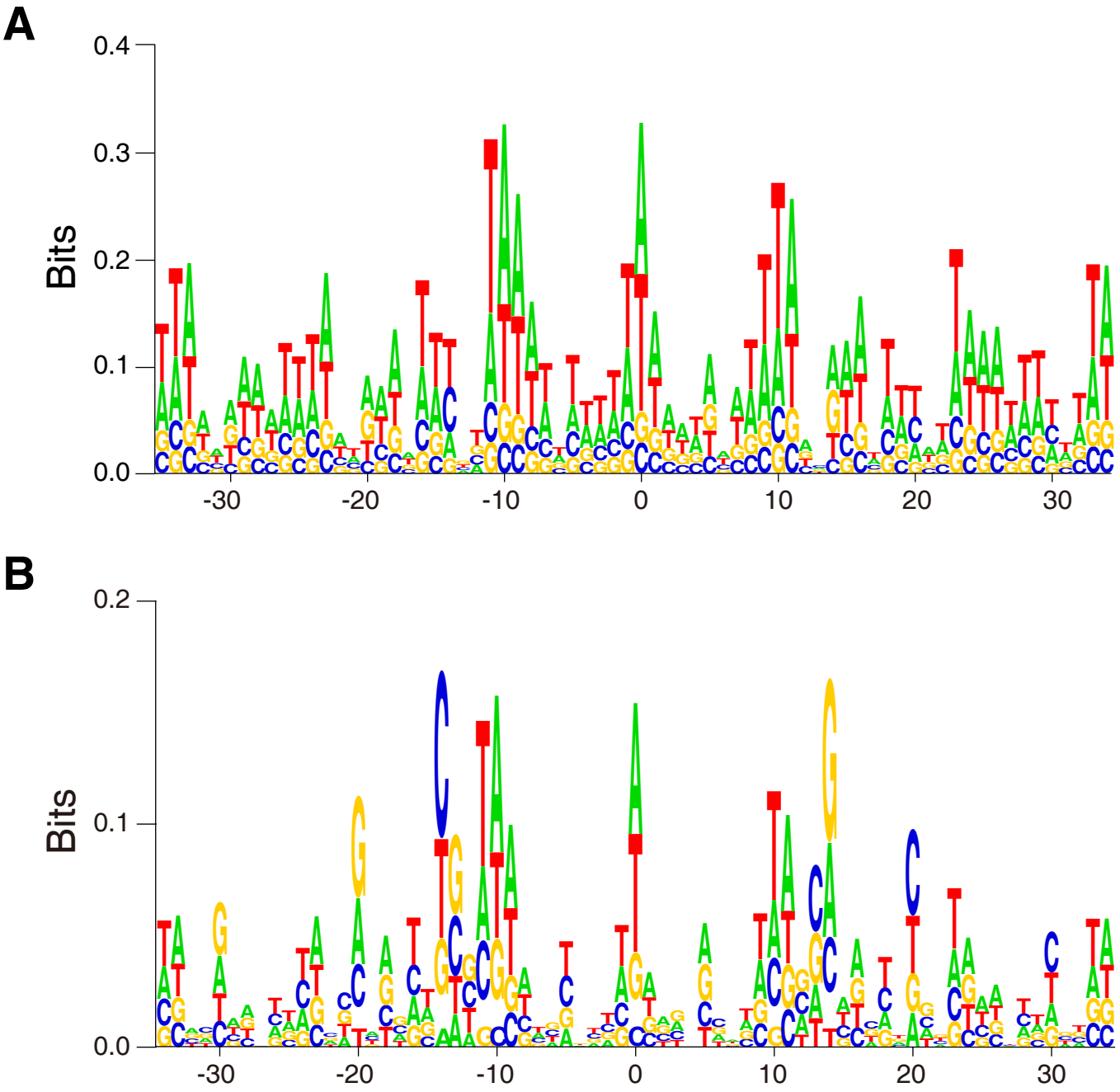
*Correspondence: guoyb9@sysu.edu.cn

Sequence logo is a powerful tool for presenting consensus sequences or motifs of nucleic acids and proteins (Schneider and Stephens 1990). WebLogo, a web-based sequence logo generator hosted by the University of California, Berkeley is the most popular logo generator so far (Crooks *et al.* 2004). WebLogo has a graphical interface and is convenient and highly configurable. However, its application is occasionally restricted by the internet speed, especially in developing countries. Moreover, when the sequence number exceeds 10,000, a command line interface will have to be used instead of graphical interface, but many users in biological sciences fields found it difficult to perform the installation and configuration of WebLogo and GhostScript (for vector map output) due to lacking relative knowledge. Here I made an application, DNAlogo, which creates DNA sequence logos in Windows with a graphical interface. DNAlogo is written in vb.net based on the algorithm and correction described previously (Schneider *et al.* 1986; Schneider and Stephens 1990).

The input for DNAlogo can be either Fasta format or pure sequences (sequence per line). DNAlogo can also read matrices designating the bit values of each character directly. The bitmaps showed in the picture box of DNAlogo can be saved as JPEG format. For publication purpose, vector maps can be output by saving PostScript (.ps) files and further processed using Adobe Illustrator.

Similar to WebLogo, DNA logo can create both bit logos and frequency logos. Frequency logos contain only the information of frequencies of the four nucleotides, but no information of sequence conservation.

Fig. 1



50,613 *S. pombe* genomic sequences flanking Tf1 integration site were aligned according to the Tf1 target site duplications and orientations, and sequence logos were created using DNAlogo. A, Regular logo; B, Compensated logo.

Because GC contents vary among genomes, DNAlogo added a new function to minimize the bias introduced by different GC contents, which is named "Compensated logo". For example, in fission yeast *Schizosaccharomyces pombe* genome, the GC content is only 36%, in which even random sequences show an AT rich pattern in sequence logos. However, random sequences should have no conservation. Fig. 1 showed the sequence pattern of the previously reported integration sites of Tf1 transposon in *S. pombe* (Guo and Levin 2010). It seems A or T is preferred in all of the positions when presented using a regular logo (Fig. 1A), but when the logo is compensated by the GC content (36%), preferences for G or C were showed in certain positions (Fig. 1B). To my knowledge, this is the first time that the GC contents were considered in sequence logo generation.

Actually, some sequence logos in two previous publications were created using DNAlogo (Guo *et al.* 2013; Chatterjee *et al.* 2014). Logos in Figure 8 of the latter are compensated logos.

DNAlogo is small, convenient, and no installation is needed. People can use DNAlogo without any programming or bioinformatics knowledge. The DNAlogo application (.exe file) and user's manual were deposited in Github (https://github.com/DNAworker/DNAlogo).

References

Chatterjee A. G., Esnault C., Guo Y., Hung S., McQueen P. G., Levin H. L., 2014 Serial number tagging reveals a prominent sequence preference of retrotransposon integration. Nucleic Acids Res. **42**: 8449–8460.

Crooks G. E., Hon G., Chandonia J. M., Brenner S. E., 2004 WebLogo: A sequence logo generator. Genome Res. **14**: 1188–1190.

Guo Y., Levin H. L., 2010 High-throughput sequencing of retrotransposon integration provides a saturated profile of target activity in Schizosaccharomyces pombe. Genome Res. **20**: 239–248.

Guo Y., Park J. M., Cui B., Humes E., Gangadharan S., Hung S., FitzGerald P. C., Hoe K. L., Grewal S. I. S., Craig N. L., Levin H. L., 2013 Integration profiling of gene function with dense maps of transposon integration. Genetics **195**: 599–609.

Schneider T. D., Stormo G. D., Gold L., Ehrenfeuch A., Ehrenfeucht A., 1986 Information content of binding sites on nucleotide sequences. J Mol Biol **188**: 415–431.

Schneider T. D., Stephens R. M., 1990 Sequence Logos : A New Way to Display Consensus Sequences. **6100**: 6097–6100.