BIOINFORMATICS APPLICATIONS NOTE Vol. 22 no. 12 2006, pages 1536–1537
doi:10.1093/bioinformatics/btl151

Sequence analysis

Two Sample Logo: a graphical representation of the differences between two sets of sequence alignments

Vladimir Vacic[1], Lilia M. Iakoucheva[2] and Predrag Radivojac[3,*]

[1]Computer Science and Engineering Department, University of California, Riverside, CA, USA,
[2]Laboratory of Statistical Genetics, The Rockefeller University, New York, NY, USA and
[3]School of Informatics, Indiana University, Bloomington, IN, USA

Received on December 29, 2005; revised on March 19, 2006; accepted on April 14, 2006
Advance Access publication April 21, 2006
Associate Editor: Martin Bishop

ABSTRACT

Sumary: Two Sample Logo is a web-based tool that detects and displays statistically significant differences in position-specific symbol compositions between two sets of multiple sequence alignments. In a typical scenario, two groups of aligned sequences will share a common motif but will differ in their functional annotation. The inclusion of the background alignment provides an appropriate underlying amino acid or nucleotide distribution and addresses intersite symbol correlations. In addition, the difference detection process is sensitive to the sizes of the aligned groups. Two Sample Logo extends WebLogo, a widely-used sequence logo generator. The source code is distributed under the MIT Open Source license agreement and is available for download free of charge.

Availability: http://www.twosamplelogo.org

Contact: predrag@indiana.edu

Bioinformatics research often requires comparative analyses of sets of sequences that differ in their functional annotation. In the case of functionally verified sequence patterns (e.g. transcription factor binding sites or protein post-translational modification sites) it may be easy to assemble a set of 'background patterns', i.e. sequences that share sequence motifs with the functionally annotated sites, but which have either different or no functional annotation. In order to visualize the differences between two such groups, we have developed Two Sample Logo, a program that generates graphical representations of statistically significant position-specific differences in amino acid or nucleotide compositions between two sets of multiply aligned sequences. Hereafter, these two sets are referred to as the positive and the negative (background) sets.

Graphical output of Two Sample Logo consists of three components: (1) an upper section displaying a set of symbols enriched (overrepresented) in the positive set; (2) a lower section displaying a set of symbols depleted (underrepresented) in the positive set; and (3) the middle section displaying consensus symbols. Symbols are organized in stacks with one stack per position in the sequence. An example of a Two Sample Logo is shown in Figure 1, where alternatively spliced exon–intron junctions are compared with the regular splice junctions.

BACKGROUND

Sequence logos were introduced by Schneider and Stephens (1990) as a way to display patterns of sequence conservation that cannot be readily seen in the outputs of standard sequence alignment programs. Crooks *et al*. (2004) subsequently developed WebLogo, a user-friendly sequence logo generator with additional features and options. Several other extensions have also been created, e.g. RNA structure logos (Gorodkin *et al*., 1997), PSSM logos (Fujii *et al*., 2004) and energy normalized sequence logos (Workman *et al*., 2005).

In its basic form, a sequence logo displays symbol information content for each position in a multiple sequence alignment. Assuming that each position in the alignment is a sample of symbols generated according to some probability distribution, the information content is calculated as the relative contribution of a symbol to the difference between maximum and estimated (observed) position-specific entropies. A known limitation of the sequence logos is that they are based on the assumptions that motif positions are mutually independent and that the same background distribution applies to every position in every motif. In addition, sequence logos are inherently insensitive to the sample size and cannot be easily used to visualize differences between two sets of alignments. Two Sample Logo offers a way to overcome these limitations through position-specific normalization using the background alignment.

STATISTICAL TESTS

For each position in the alignment and each symbol in the alphabet, the program first assembles binary vector representations of symbol incidence. Using either two sample *t*-test or binomial test, Two Sample Logo then evaluates the hypothesis that the vector from the positive set and the corresponding vector from the negative set were generated by the same distribution.

The two supported statistical tests are based on different underlying assumptions. In particular, the two sample *t*-test assumes that the samples are normally distributed with equal variances and the null hypothesis is that the means of the samples are identical. This test is computationally fast and is known to be robust to the violation of the normality assumption. The binomial test is based on the assumption that an occurrence of a symbol at any position follows the binomial distribution. It estimates the significance level of the hypothesis that symbol occurrence probabilities are identical in both samples. A more detailed explanation of the statistical tests is provided in the online documentation.

[*]To whom correspondence should be addressed at School of Informatics, Indiana University, 1900 East 10th Street, Eigenmann Hall 1005, Bloomington, IN 47406, USA

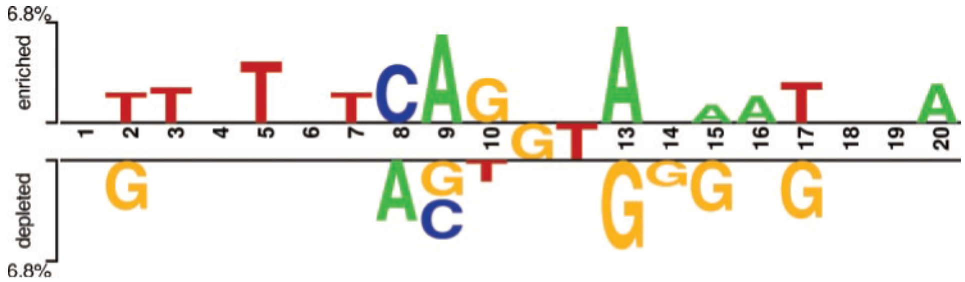

Figure 1. Two Sample Logo of the differences between 2000 alternatively and 2000 regularly spliced GT exon–intron junctions for the significance threshold of 0.05. Alternatively spliced sites (positive set) were extracted from HASDB (Modrek, *et al*., 2001) as 20 nt-long sequences around 5′ splice sites, centered around a GT dinucleotide, which had more than one competing 3′ site. Regular splice sites (negative set) were taken from a set of all non-identical exon–intron junctions from HS^{3}D (Pollastro and Rampone, 2002). Both alternatively and regularly spliced sites were selected as random samples from the corresponding repositories.

Both tests reduce the number of potentially displayed symbols by presenting only statistically significant subsets (the *P*-value threshold is a user-specified parameter). In addition, there is an option to use Bonferroni correction to eliminate spurious statistical significance. Both statistical tests assume that each sequence in each alignment is independent of others. If closely homologous sequences are present in the datasets, a length-dependent scheme to remove redundancy is recommended (Rost, 1999).

GRAPHICAL OUTPUT

Two Sample Logo produces two variants of graphical output, i.e. statistically significant symbols may be displayed using fixed or variable heights. Variable symbol heights are proportional to the difference in symbol frequency between the samples. The software provides a number of pre-defined color schemes that assist in visual identification of standard physicochemical properties and also supports user-defined color schemes.

ACKNOWLEDGEMENTS

Both web and command line versions of Two Sample Logo were written in Ruby and extend the freely available WebLogo code. Computation-intensive routines for calculating *P*-values were written in C and use numerical approximation functions from the Stephen L. Moshier's Cephes Math Library, available for download from http://www.netlib.org/cephes. The authors thank Mehmet M. Dalkilic for proofreading the manuscript.

Conflict of Interest: none declared.

REFERENCES

Crooks,G.E. *et al*. (2004) WebLogo: a sequence logo generator. *Genome Res*., **14**, 1188–1190.

Fujii,K. *et al*. (2004) Kinase peptide specificity: improved determination and relevance to protein phosphorylation. *Proc. Natl Acad. Sci. USA*, **101**, 13744–13749.

Gorodkin,J. *et al*. (1997) Displaying the information contents of structural RNA alignments: the structure logos. *Comput. Appl. Biosci*., **13**, 583–586.

Modrek,B. *et al*. (2001) Genome-wide detection of alternative splicing in expressed sequences of human genes. *Nucleic Acids Res*., **29**, 2850–2859.

Pollastro,P. and Rampone,S. (2002) HS^{3}D, a dataset of *Homo sapiens* splice regions and its extraction procedure from a major public database. *Int. J. Mod. Phys*., **13**, 1105–1117.

Rost,B. (1999) Twilight zone of protein sequence alignments. *Protein Eng*., **12**, 85–94.

Schneider,T.D. and Stephens,R.M. (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res*., **18**, 6097–6100.

Workman,C.T. *et al*. (2005) enoLOGOS: a versatile web tool for energy normalized sequence logos. *Nucleic Acids Res*., **33**, W389–W392.