

Two Sample Logo User Manual

Vladimir Vacic* Lilia M. Iakoucheva† Predrag Radivojac‡

Version 1.2

1 Introduction

Two Sample Logo is a procedure for discovery of statistically significant position-specific differences in residue compositions between two multiple sequence alignments, as well as for graphical representation of those differences. It was proposed by Vacic, Iakoucheva and Radivojac [11], and implemented as a web application (available at <http://www.twosamplelogo.org>) and as a command line program. Both versions of the software were written in Ruby and extend the freely available WebLogo code [2].

2 Installation

2.1 Download

Two Sample Logo can be downloaded from <http://www.twosamplelogo.org>.

2.2 System requirements

Command line version of Two Sample Logo requires that you have a Ruby interpreter installed on your system, as well as GhostScript (a PostScript interpreter) and ImageMagick. All three programs are by default installed on any Linux system; in the event that they are not, they can be downloaded free of charge from:

Ruby - <http://www.ruby-lang.org>

GhostScript - <http://www.cs.wisc.edu/~ghost>

*Department of Computer Science and Engineering, University of California, Riverside, CA

†Laboratory of Statistical Genetics, The Rockefeller University, New York, NY

‡School of Informatics, Indiana University, Bloomington, IN

ImageMagick - <http://www.imagemagick.org>

The Ruby interpreter, GhostScript and ImageMagick are by default in the system path. In the case that they are not, the paths to the appropriate executables can be specified in the `ts1.conf` file, for example:

```
gs=/usr/bin/gs
convert=/usr/bin/convert
```

In addition, the web version requires a running web server: Two Sample Logo has been tested on Apache, using the Ruby module.

2.3 Compilation

Most of the Two Sample Logo code is written in Ruby and does not need to be compiled. The only exception are the routines for computing the statistical significance, which were coded in C due to efficiency concerns.

The statistical significance code can be found in the `pvalue` directory. Before it can be used, it needs to be compiled on the computer on which it will be run. A `Makefile` is provided; it is sufficient to type `make` on the command prompt in the `pvalue` directory. The resulting `pvalue` executable should be copied to the directory with the Ruby scripts (e.g. `cgi-bin`).

2.4 Web application

To improve security, we have separated the CGI scripts from the HTML documents and images. Cgi scripts are in a subdirectory of `cgi-bin`, and HTML documents are in a subdirectory of the Apache web document root. On Apache web server running on a Linux system, assuming the default Apache settings, those are `/var/www/cgi-bin/ts1` and `/var/www/html/ts1`, respectively.

For the Two Sample Logo CGI scripts to be able to link to HTML documents (such as help files, etc.), relative path for the HTML documents in relation to the CGI scripts directory has to be specified in `ts1.cgi` "path" variable. For example:

```
path = "../../../ts1/"
```

In addition to this, the `ts1.cgi` script needs to be configured to write the output images into an Apache-writable directory under the Apache web document root, so they can be displayed to the user. This is done using the "temp" variable. For example:

```
temp = "/var/www/html/tsl/cache/"
```

2.5 Licensing

Two Sample Logo is distributed under the MIT Open Source License (<http://www.opensource.org/licenses/mit-license.html>). A full license document can be found in the on-line license document: <http://www.twosamplelogo.org/LICENSE.txt>.

3 Running Two Sample Logo

3.1 Command line version

Even though the web version of Two Sample Logo is sufficient for both a preliminary data analysis and generating publication-quality figures, the command line version of Two Sample Logo allows more flexibility for the users. The command line version is a Ruby script called `tsl`. For a complete list of options, `tsl` can be invoked with the `-h` switch:

```
Usage: tsl -P <pos file> -N <neg file> -K[A|N] -O <out file> [options]
Creates a two sample logo from the two multiple sequence alignments.
```

Mandatory arguments:

<code>-P <positive sample file></code>	
<code>-N <negative sample file></code>	
<code>-K <kind of data></code>	A for amino acid, N for nucleic acid.
<code>-O <output file></code>	Output file name.

Optional arguments:

<code>-A <box shrink factor></code>	Shrink factor of characters if option <code>s</code> (show box) is toggled. Defaults to 0.5.
<code>-C <color scheme></code>	One of the following: amino_bw (Black and white) amino_weblogo (WebLogo default colors) amino_colors (Amino colors) amino_shapley (Shapley colors) amino_charge (Charge) amino_hydro (Hydrophobicity) amino_surface (Surface exposure) amino_flex (Vihinen's flexibility) amino_disorder (Disorder propensity) nucleo_bw (Black and white) nucleo_weblogo (WebLogo default colors) nucleo_shapley (Shapley colors)

	Defaults to amino_weblogo.
-E <sequence end>	
-F <format>	Format of output (EPS, GIF, PDF, PNG, TXT). Defaults to PNG.
-H <logo height>	Height of output logo. Defaults to 5.
-I <title>	Text of title, enclosed in "" if more than one word.
-M <first index>	Defaults to 1.
-R <resolution>	Bitmap resolution. Defaults to 96.
-S <sequence start>	Sequence start number. Defaults to 1.
-T <statistical test>	"ttest" for t-test, "binomial" for binomial test. Defaults to "ttest".
-U <units>	Logo dimensions units (cm, inch, pixel, point). Defaults to cm.
-V <p value>	
-W <logo width>	Width of output logo. Defaults to 8.
-X <res units>	Resolution units when bitmap resolution is specified (ppi, ppc, ppp). Defaults to ppi.

Optional toggles (no values associated):

-a	Toggle anti-aliasing.
-b	Toggle Bonferroni correction.
-f	Toggle fixed height output characters.
-o	Toggle outlining of characters.
-s	Toggle show box.
-x	Toggle numbering along the x-axis.
-y	Toggle labels on y-axis.

3.2 Graphical output

Graphical output of the Two Sample Logo method 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. Symbol heights are proportional to the difference in symbol frequency between the samples.

An example of a Two Sample Logo is shown in Figure 1 C), which gives the representation of the ± 12 neighborhoods of ubiquitinated lysines. Figure 1 A) and B) gives sequence logos for comparison: A) is the default 4-bit magnification which shows that no amino acid conservation signal is strong enough, and B) is the same sequence logo magnified to 0.4-bit, at which resolution the picture is too noisy to be meaningful.

3.3 Textual output

The `-O TXT` option outputs the underlying p-values as a plain text file. Filtering of the output will be based on the chosen statistical test (see Section 5 for technical details) and the p-value cutoff. To see all raw p-values, the cutoff should be set to 1 (`-V 1`) or higher if Bonferroni correction is used.

Note that for positions which contain conserved residues there is no difference between the positive and negative sample and such positions will be skipped in the textual output.

3.4 Command line example

Using the sequence alignments provided in the Two Sample Logo distribution as positive and negative samples, a sample TSL can be generated using the following command line options:

```
./tsl -P ../datasets/phos_Y_pos.txt -N ../datasets/phos_Y_neg.txt -K A -F PNG \
-O testing.png -x -y -C amino_charge -M -12 -a
```

4 Background

Sequence logos were introduced by Schneider and Stephens [10] 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.* [2] subsequently developed WebLogo, a user-friendly sequence logo generator with additional features and options.

A basic form of 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 from some alphabet \mathcal{A} , generated according to some probability distribution, the information content is calculated as the relative contribution of a symbol to the difference between S_{max} , the maximum entropy, and S_{obs} , the estimated (observed) position-specific entropy:

$$R_{seq} = S_{max} - S_{obs} = \log_2 |\mathcal{A}| + \sum_{a \in \mathcal{A}} p_a \cdot \log_2 p_a$$

where p_a denotes the observed frequency of symbol a at a particular position in the sequence.

Sequence logos implicitly assume that motif positions are mutually independent and that the same background distribution applies to each position in every motif. They are also inherently insensitive to the sample size, e.g., if observed frequency of symbol a at a particular sequence position is 0.6 sequence logos cannot distinguish between seeing 3 occurrences out of 5, or 300 out of 500. Finally, sequence logos cannot be easily used to visualize differences between two sets of alignments. Two Sample Logo (TSL) overcomes these limitations through position-specific normalization using the background alignment.

5 Technical details

For each position in the sequence alignment, Two Sample Logo calculates statistical significance of the differences in symbol occurrence frequencies between two sets of aligned sequences, and in order to ease interpretation of results, reports only those differences which fall below the statistical significance threshold. Sequences that contain the motif and have a certain functional property constitute a *positive sample*. Sequences that contain the motif but at the same time do not have the functional property or have a contrasting functional property constitute the *negative sample*. We note that, strictly speaking, the distinction between the samples does not necessarily have to be based on the presence and absence of a functional property: as long as there is a clear way of interpreting the data, any pair of sets of sequence alignments can be used as a positive and negative set.

More formally, let P and Q be two sequence alignments based on the positive and the negative sample respectively, and let $|P|$ and $|Q|$ denote the numbers of sequences in these alignments. Let N be the length of each sequence in the two alignments (note that we require that all sequences have the same length). Let P_i denote the i^{th} sequence in alignment P , and let $P_{i,j}$ denote the j^{th} position in sequence P_i . For each of the N positions in the alignments and for each symbol a from the alphabet \mathcal{A} , we form binary vectors $X_P^{j,a}$ of indicator variables $I(P_{i,j} = a)$, $i = 1, 2, \dots, |P|$ which take a value 1 if condition t of $I(t)$ is satisfied and 0 otherwise. Vector $X_Q^{j,a}$ is conversely formed. We then estimate the p-value of the *null* hypothesis that vectors $X_P^{j,a}$ and $X_Q^{j,a}$ were sampled from the same distribution, that is, that occurrence probabilities for symbol a are identical at position j in both samples. We note that it follows from the construction that $|X_P| = |P|$ and $|X_Q| = |Q|$.

P-value is defined as the lowest significance level at which the *null* hypothesis can be rejected; it is calculated as the probability that the test statistic T equal or more extreme than θ , the observed value of the test statistic, can occur by chance alone if we assume that the *null* hypothesis holds. For the purposes of hypothesis testing, the p-value is compared with the significance level α , and in the event that p-value is less than α , the *null* hypothesis is rejected.

We here use either the two sample t-test or binomial test to estimate the p-value. These tests are based on different underlying assumptions and have slightly different properties.

5.1 Two sample t-test

The two sample t-test assumes that the occurrences of symbol a at position j in the alignments are normally distributed [5]. The *null* hypothesis is that the two Gaussians that X_P and X_Q have been drawn from have equal means:

$$H_0 : \mu_P = \mu_Q$$

This test is computationally fast and is known to be robust to the violation of the normality assumption, both of which make it a frequently used statistical procedure. We employ the unpaired version of the test, because the positive and the negative samples are independent of each other.

5.2 Binomial test

The binomial test is based on the assumption that an occurrence of a given symbol at any fixed position in the alignment follows the binomial distribution. The *null* hypothesis is that p , the probability of success parameter, is the same in both distributions which the positive and negative samples have been drawn from:

$$H_0 : p_P = p_Q$$

The test statistic is the absolute difference in the relative frequencies of the symbol occurrence, which are the maximum likelihood estimates for the probability of success:

$$T = \left| \frac{k_P}{|P|} - \frac{k_Q}{|Q|} \right|$$

where k_P and k_Q are the numbers of successful trials, that is, numbers of ones in the binary vectors X_P and X_Q . Let θ be the observed value of the test statistic T . We determine the critical region of the test statistic by directly computing the probability of $P(T \geq \theta)$ event occurring by chance alone. According to the *null* hypothesis, the maximum likelihood estimate for the parameter p of the binomial distribution is the relative frequency of observing the symbol when the vectors X_P and X_Q are concatenated:

$$p = \frac{k_P + k_Q}{|P| + |Q|}$$

and the achieved significance level of the *null* hypothesis is:

$$p - value = \sum_{\substack{0 \leq i \leq |P| \\ 0 \leq j \leq |Q| \\ \left| \frac{i}{|P|} - \frac{j}{|Q|} \right| \geq \theta}} \binom{|P|}{i} \cdot p^i (1-p)^{|P|-i} \cdot \binom{|Q|}{j} \cdot p^j \cdot (1-p)^{|Q|-j}$$

5.3 Multiple test correction

When a number of statistical tests are performed simultaneously, there arises a chance that some of the computed p-values are going to be below the threshold α due to chance alone. That is, while α is appropriate for each individual statistical test, it may not be appropriate for a set of n tests. The simplest and most conservative multiple testing correction procedure which removes some of the spurious significance is due to Bonferroni [1]. Here the α value is divided by n , the number of tests performed.

The Bonferroni correction follows from the Bonferroni inequality:

$$P \left(\bigcup_{i=1}^n E_i \right) \leq \sum_{i=1}^n P(E_i)$$

which states that a probability of a union of events (probability that at least one event will be observed) is bounded by the sum of probabilities of individual events. The equality is achieved when the events are statistically independent.

In order not to bias the discovered motifs towards the motifs which occur in closely homologous sequences, if such sequences are present in the datasets, a length-dependent scheme for removing redundancy due to Rost [9] is recommended.

6 Examples

6.1 Ubiquitination

Ubiquitination is a protein modification in which ubiquitin molecule covalently binds to the lysine residues of ubiquitin substrates [3]. This two sample logo example contains 25 residue wide fragments, 12 upstream and 12 downstream, from all lysines found in the 95 ubiquitinated proteins reported in [4, 7]. The positive sample contains 110 non-redundant fragments around experimentally verified ubiquitination sites, while the negative sample contains all remaining lysines from the same set of proteins, 2885 in total. In order to help answering a question about sequence biases around ubiquitination sites, a two sample logo can be generated to visualize residues that are significantly enriched or depleted in the set of ubiquitinated fragments.

6.2 Calmodulin IQ motifs

Calmodulin signaling involves important and wide-spread eukaryotic protein-protein interactions that regulate a variety of cellular processes [12]. Among five major types of calmodulin binding sites, one particular type of calmodulin regulation is predominantly calcium independent and requires so-called IQ motifs as the calmodulin binding regions. This two sample logo example is based on the Calmodulin Target Database [13] and contains a set of 17 non-redundant experimentally verified IQ motifs and 294 remaining fragments, all containing dipeptide IQ, from the same set of proteins. A two sample logo can be generated in order to understand sequence biases that distinguish IQ motifs from the remaining fragments that contain IQ, but are not calmodulin-binding.

Calmodulin Target Database suggests that a regular expression for the calmodulin IQ motif is `[FILV]QXXX[RK]GXXX[RK]XX[FILVWY]`. However, even though positive and negative sequences in this example were restricted to only those starting with an IQ dipeptide, the two sample logo reveals interesting dependencies when compared to the negative sites.

6.3 Exon-intron splice sites

This two sample logo displays the differences between experimentally verified and false positive exon-intron junction sites, both downloaded from the HS³D database [8]. The positive set contains 2,000 non-identical 20-nucleotide long exon-intron junctions centered at GT

dinucleotide. The negative set contains 2,000 non-identical randomly selected false positive junctions, also centered at GT. In contrast to a one-sample WebLogo output, this example clearly indicates higher GC content on the intron (3') side of the junction.

Interestingly, both 5' and 3' sides show significant depletion of thymine, except at position 15 (6th base within an intron), where it is significantly enriched in the positive dataset.

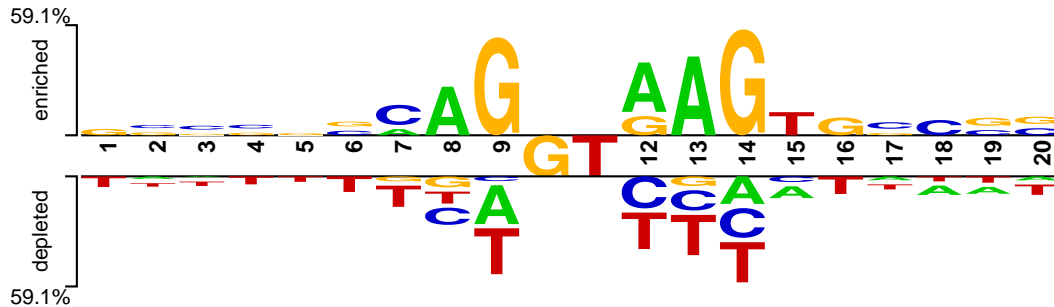


Figure 3: Two sample logo for the splicing junctions.

6.4 Alternatively and regularly spliced exon-intron junctions

Two sample logo of the differences between alternatively and regularly spliced exon-intron junctions for the p-value threshold of 0.05. A random sample of 2,000 alternatively spliced sites centered at GT dinucleotide (positive sample) was extracted from HASDB [6] as 20 nucleotide-long sequences around 5 splice sites which had more than one competing 3 site. Regular splice sites (negative sample) consisted of a random sample of 2,000 non-identical exon-intron junctions from HS3D database [8].

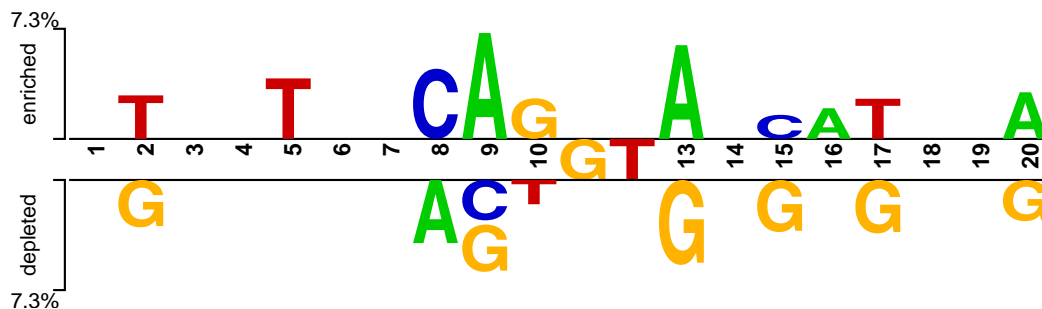


Figure 4: Two sample logo for the difference between alternatively and regularly spliced exon-intron junctions.

7 Acknowledgments

Two Sample Logo is derived from WebLogo, a sequence logo generator written by Crooks *et al.* [2] and available on-line at <http://weblogo.berkeley.edu>.

Numerical approximation functions are taken from Stephen L. Moshier's *Cephes Math Library* (available on-line at <http://www.netlib.org/cephes>) and incorporated herein by permission of the author.

References

- [1] C. E. Bonferroni. Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, 8:3–62, 1936.
- [2] G. E. Crooks, G. Hon, J. M. Chandonia, and S. E. Brenner. WebLogo: a sequence logo generator. *Genome Research*, 14:1188–90, 2004.
- [3] L. Hicke. Protein regulation by monoubiquitin. *Nature Reviews Molecular Cell Biology*, 2:195–201, 2001.
- [4] A. L. Hitchcock, K. Auld, S. P. Gygi, and P. A. Silver. A subset of membrane-associated proteins is ubiquitinated in response to mutations in the endoplasmic reticulum degradation machinery. *Proceedings of the National Academy of Sciences*, 100:12735–40, 2003.
- [5] R. V. Hogg, A. Craig, and J. W. McKean. *Introduction to mathematical statistics*. Prentice Hall, Upper Saddle River, NJ, 6th edition, 2004.
- [6] B. Modrek, A. Resch, C. Grasso, and C. Lee. Genome-wide detection of alternative splicing in expressed sequences of human genes. *Nucleic Acids Research*, 29(13):2850–9, 2001.
- [7] J. Peng, D. Schwartz, J. E. Elias, C. C. Thoreen, D. Cheng, G. Marsischky, J. Roelofs, D. Finley, and S. P. Gygi. A proteomics approach to understanding protein ubiquitination. *Nature Biotechnology*, 21:921–6, 2003.
- [8] P. Pollastro and S. Rampone. HS3D, a dataset of *Homo sapiens* splice regions and its extraction procedure from a major public database. *International Journal of Modern Physics*, 13:1105–17, 2002.
- [9] B. Rost. Twilight zone of protein sequence alignments. *Protein Engineering*, 12:85–94, 1999.
- [10] T. D. Schneider and R. M. Stephens. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Research*, 18:6097–100, 1990.

- [11] V. Vacic, L. M. Iakoucheva, and P. Radivojac. Two Sample Logo: a graphical representation of the differences between two sets of sequence alignments. *Bioinformatics*, 22(12):1536–7, 2006.
- [12] S. W. Vetter and E. Leclerc. Novel aspects of calmodulin target recognition and activation. *FEBS Journal*, 270(3):404–14, 2003.
- [13] K.L. Yap, J. Kim, K. Truong, M. Sherman, T. Yuan, and M. Ikura. Calmodulin target database. *Journal of Structural and Functional Genomics*, 1(1):8–14, 2000.