pLogo: a probabilistic approach to visualizing sequence motifs

Joseph P O'Shea[1,3], Michael F Chou[2,3], Saad A Quader[1], James K Ryan[1], George M Church[2] & Daniel Schwartz[1]

Methods for visualizing protein or nucleic acid motifs have traditionally relied upon residue frequencies to graphically scale character heights. We describe the pLogo, a motif visualization in which residue heights are scaled relative to their statistical significance. A pLogo generation tool is publicly available at http://plogo.uconn.edu/ and supports real-time conditional probability calculations and visualizations.

Sequence motifs, short linear patterns of nucleotides or amino acids, are known to serve as molecular beacons for a wide variety of biological processes[1–6]. Consensus-derived representations of biological motifs are frequently oversimplified models of binding specificities because instances of functional motifs often do not conform to the consensus. A number of tools and methodologies have, over the last two decades, provided improved visual representations of biological motifs[7–10]. These methodologies differ from each other in the metric by which individual residues are scaled and in the background model assumed. Here we introduce the probability logo (pLogo) to address these two facets of motif visualization.

A pLogo image has the following features. (i) Residues are scaled relative to their statistical significance (rather than to their frequency) in the context of proteomic, genomic or user-defined backgrounds. (ii) There is a depiction of over- and underrepresented residues, which are shown above and below the *x* axis, respectively. (iii) Horizontal ruler lines above and below the *x* axis correspond to Bonferroni-corrected statistical significance values (these significance lines serve as a useful reference across different motifs). (iv) Residues are stacked in order of statistical significance, with the most significant residues positioned closest to the *x* axis (this positioning allows one to easily read the consensus sequence just along the top of the axis, and it also allows direct visual comparison of the significance ruler lines with the height of the most significant residues). (v) A key at the bottom designates the statistical significance thresholds as well as the number of aligned sequences in the input data set (which we refer to as the 'foreground') and the background data set that were used to generate the image. (vi) 'Fixed' motif positions are highlighted.

The ability to dynamically 'fix' and 'unfix' motif positions is provided by the pLogo Web tool, thus creating a real-time conditional probability framework to examine biological sequence motifs. This framework is a direct extension of the widely used automated motif-extraction algorithm we previously developed

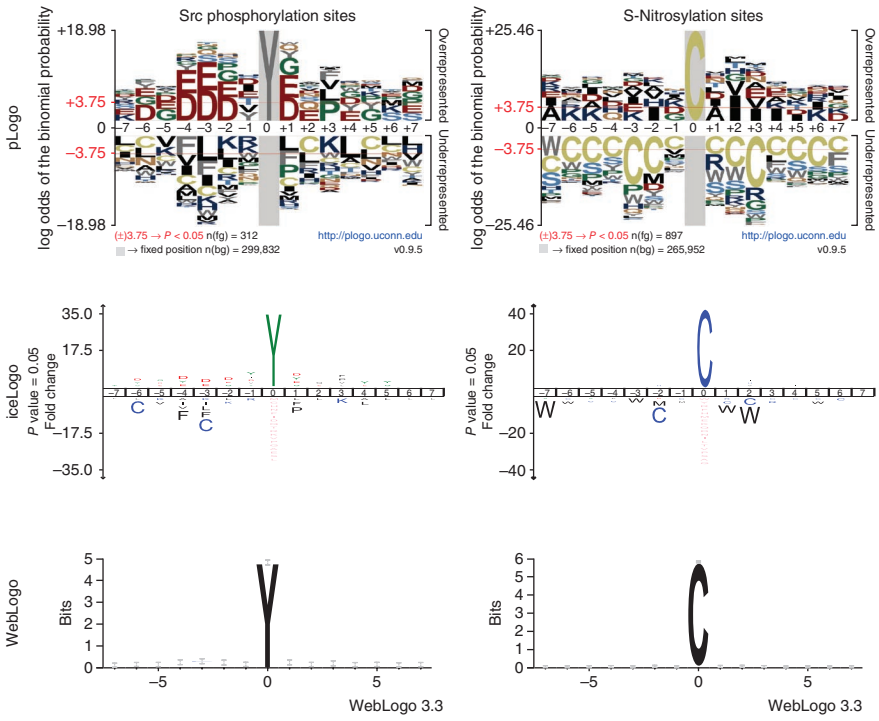

Figure 1 | Comparison of sequence logo visualizations for the pLogo, iceLogo and WebLogo tools. 312 human Src tyrosine phosphorylation sites (left) and 897 mouse S-nitrosylated cysteine sites (right) were used as input for each visualization strategy. Images are scaled to the height of the largest column within the sequence visualization. The n(fg) and n(bg) values at the bottom left of the pLogo indicate the number of aligned foreground and background sequences used to generate the image, respectively. The red horizontal bars on the pLogo correspond to $P = 0.05$.

[1]Department of Physiology and Neurobiology, University of Connecticut, Storrs, Connecticut, USA. [2]Department of Genetics, Harvard Medical School, Boston, Massachusetts, USA. [3]These authors contributed equally to this work. Correspondence should be addressed to D.S. (daniel.schwartz@uconn.edu).

RECEIVED 4 APRIL; ACCEPTED 19 AUGUST; PUBLISHED ONLINE 6 OCTOBER 2013; DOI:10.1038/NMETH.2646

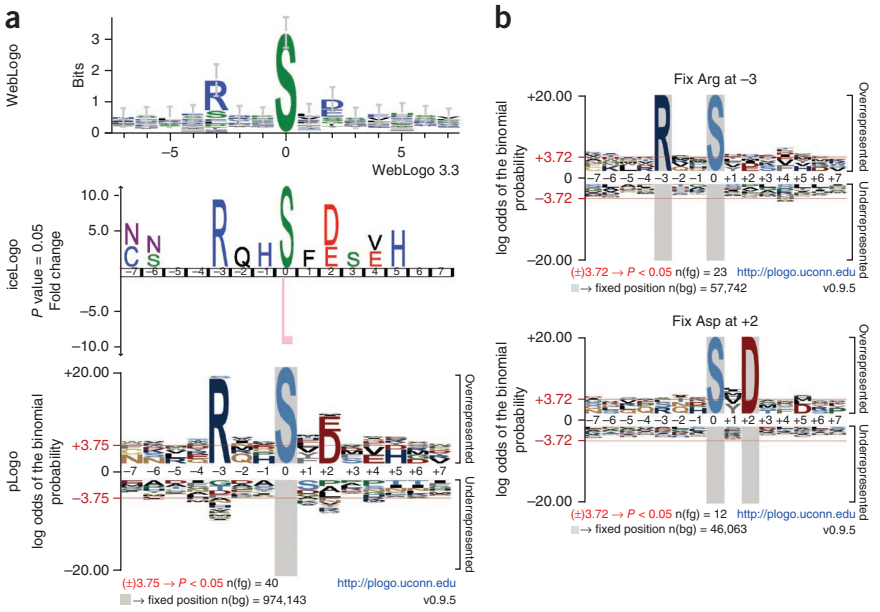

Figure 2 | Demonstration of the use of pLogo conditional probabilities through fixed positions. (**a**) Known sites from human CaMKII substrates (n = 40 sites) were used to generate motif visualizations from the indicated tools. (**b**) Two residues were independently fixed in the pLogo, and these modifications reveal new pLogos in the context of the fixed positions.

known as motif-x[11]. As with motif-x, pLogos use binomial probabilities to assess the statistical significance of motif residues at given positions[12–14].

We compared two post-translational modification (PTM) data sets, Src kinase substrates in the human proteome (n = 312 sites) and S-nitrosylation sites in the mouse proteome (n = 897 sites), using the pLogo, iceLogo[9], Two Sample Logo[8] and WebLogo[15] tools (**Fig. 1**, **Supplementary Fig. 1** and **Supplementary Table 1**). The same foreground data set was uploaded to each tool, but the background data sets differed according to the expected usage of the tool in question. In a pLogo, the probability-based scaling of residue heights, and the decoupling of the central residue from the scale of the y axis, yielded information-rich visualizations.

Because motif visualizations are a representation of overall linear protein (or DNA) sequence data, any intraresidue correlations are lost when the other visualization tools are used. We overcome this limitation with the pLogo generation tool by providing a mechanism to fix/unfix residues in real time. The resulting pLogos thus exhibit the conditional probabilities of unfixed residues given the fixed residues in the context of a similarly fixed background, as shown for the calmodulin-dependent protein kinase II (CaMKII) motif obtained from 40 nonredundant CaMKII phosphorylation sites[16] (**Fig. 2a**, **Supplementary Fig. 2** and **Supplementary Table 2**). WebLogo, iceLogo and Two Sample Logo implied an overall CaMKII motif of the general form RXXSXD; however, by using the pLogo generation tool to independently fix the arginine at –3 and the aspartic acid at +2, we could readily observe the lack of correlation between these two otherwise significant residues (**Fig. 2b**). This finding is consistent with literature documenting that CaMKII is able to independently phosphorylate both RXXS and SXD sequences[17]. The ability to fix residues and in turn detect conditional probabilities is important in PTM data sets, for instance, in which multiple enzymes with potentially different sets of specificities might act on a set of substrates (**Supplementary Figs. 3** and **4** and **Supplementary Tables 3** and **4**). pLogo visualizations are equally effective in the representation of DNA motifs (**Supplementary Fig. 5** and **Supplementary Table 5**).

As genomic and proteomic sequencing data continue to grow, so does the need for tools that can accurately summarize patterns within data in a maximally informative and biologically consistent manner. Beyond visualization, the underlying probability-based framework described here may find applications that guide the next generation of predictive software toward the ultimate goal of improving hypothesis generation, experimentation and discovery.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

The authors wish to thank J. Lubner for his assistance in beta-testing the pLogo Web tool. Additionally, we thank the University of Connecticut Bioinformatics Facility for hosting the pLogo website and maintaining the server on which it runs. This study was funded in part by grants from the University of Connecticut Research Foundation (D.S.) and a Genomes to Life grant from the US Department of Energy (G.M.C.).

AUTHOR CONTRIBUTIONS

M.F.C. and D.S. conceived of and designed the study. J.P.O., M.F.C. and D.S. gathered and analyzed the data. J.P.O. and D.S. designed the website. J.P.O., S.A.Q. and J.K.R. coded the website. G.M.C. provided materials and experimental insights. D.S. wrote the initial manuscript. All authors helped edit the final manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Portales-Casamar, E. *et al. Nucleic Acids Res.* **38**, D105–D110 (2010).
2. Catterall, J.F. *et al. Nature* **275**, 510–513 (1978).
3. Munro, S. & Pelham, H.R. *Cell* **48**, 899–907 (1987).
4. Miller, M.L. *et al. Sci. Signal.* **1**, ra2 (2008).
5. Dahiya, A., Gavin, M.R., Luo, R.X. & Dean, D.C. *Mol. Cell Biol.* **20**, 6799–6805 (2000).
6. Saraste, M., Sibbald, P.R. & Wittinghofer, A. *Trends Biochem. Sci.* **15**, 430–434 (1990).
7. Schneider, T.D. & Stephens, R.M. *Nucleic Acids Res.* **18**, 6097–6100 (1990).
8. Vacic, V., Iakoucheva, L.M. & Radivojac, P. *Bioinformatics* **22**, 1536–1537 (2006).
9. Colaert, N., Helsens, K., Martens, L., Vandekerckhove, J. & Gevaert, K. *Nat. Methods* **6**, 786–787 (2009).
10. Workman, C.T. *et al. Nucleic Acids Res.* **33**, W389–W392 (2005).
11. Schwartz, D. & Gygi, S.P. *Nat. Biotechnol.* **23**, 1391–1398 (2005).
12. Prisic, S. *et al. Proc. Natl. Acad. Sci. USA* **107**, 7521–7526 (2010).
13. Chiang, C.W. *et al. Genetics* **180**, 2277–2293 (2008).
14. Chou, M.F. *et al. PLoS ONE* **7**, e52747 (2012).
15. Crooks, G.E., Hon, G., Chandonia, J.M. & Brenner, S.E. *Genome Res.* **14**, 1188–1190 (2004).
16. Hornbeck, P.V. *et al. Nucleic Acids Res.* **40**, D261–D270 (2012).
17. Feinmesser, R.L., Wicks, S.J., Taverner, C.J. & Chantry, A. *J. Biol. Chem.* **274**, 16168–16173 (1999).

ONLINE METHODS

Position weight matrix (PWM) scores used for pLogo generation. The pLogo is a visual representation of a PWM in which each possible residue has a value at each position. These values are based on the statistical significance of the frequency of that residue in the foreground data set given the probability of that residue in the background data set. Thus, a 15-base-pair DNA motif would have a total of 15 positions × 4 nucleotides = 60 values, and a 15-residue protein motif would have 15 positions × 20 amino acids = 300 values. In a pLogo, each of these values is visually represented by the height of the particular residue either above or below the x axis depending on the sign (+ or −) of the value. Each residue value at each position is calculated as described next.

To determine one-tailed significance, one normally computes the area under the desired side of a probability distribution. Using this method, one needs to use logic and cases to determine the appropriate side of the distribution to integrate (i.e., overrepresentation versus underrepresentation). For pLogo generation, however, a continuous probability-based score spanning underrepresentation, overrepresentation and values in between was desirable. Thus, we chose to make residue heights in a pLogo closely proportional to the log odds of the significance of overrepresentation versus the significance of underrepresentation. These significance values are calculated using the binomial probability of residue frequencies, with respect to a genomic or proteomic background, using the following formula:

$$\text{Residue height}(K,N,p) \propto -\log\frac{\Pr(k,\forall k \geq K \mid N,p)}{\Pr(k,\forall k \leq K \mid N,p)}$$

where K is the actual number of residues of a particular type at a given position, N is the total number of residues at a position and p is the probability of the residue occurring at that position. The value of p is derived from the background population (see "Calculation of background probabilities"), such that

$$\Pr(k,\forall k \geq K,N,p) = \sum_{k=K}^{N} \text{binomial}(k,N,p)$$

$$\Pr(k,\forall k \leq K,N,p) = \sum_{k=0}^{K} \text{binomial}(k,N,p)$$

Note that in this formulation, the probability of exactly K occurrences, $\Pr(K, N, p)$, is actually double counted because it is included in both tails of the distribution. Therefore, because the two tails will sum to slightly greater than 1.0, this is not strictly the log odds but is instead a very close approximation. Additionally, as N increases, residue heights approach the exact log-odds value because $\Pr(K, N, p)$ diminishes toward 0. In the case of $\lfloor N \times p \rfloor \leq K \leq \lceil N \times p \rceil$, this log-odds approximation will be near 0, which is the desired result when residues are neither underrepresented nor overrepresented.

The use of this log-odds approximation of the binomial probability for pLogo residue heights has the following attractive properties: (i) it is intuitive, as it closely represents the log odds of overrepresentation; (ii) it has a value of 0 (or very close to 0) when a residue is neither underrepresented nor overrepresented; (iii) it is well defined for every value of K from $0 \leq K \leq N$; (iv) it is monotonic and passes near 0 as K approaches the value of $N \times p$; and (v) it quickly approximates the true log odds and also $-\log(\Pr(k, \forall k \geq K \mid N, p))$ or $-\log(\Pr(k, \forall k \leq K \mid N, p))$ as K is farther from $N \times p$. (Therefore, if a residue is substantially underrepresented or overrepresented, one probability will be very small and the other will be near 1.0; thus, being toward one tail or the other will hardly be affected at all by the weight of the other tail).

All logs are calculated in base 10.

Calculation of background probabilities. For each position of a pLogo, the residue height is calculated according to the probability of that same residue at that same position in the background data set (or subset of the background data set if any residues are fixed at one or more positions). The model assumes a statistically large background (usually genomic or proteomic in scale), and thus the probability p is calculated as the fraction of that residue at that position in the background.

Fixing motif residues. The procedure for fixing motif residues simply involves using a subset of each of the foreground (i.e., the sequences within which one is trying to find a motif) and the background data sets to contain only those sequences bearing the desired residue at the fixed position. Subsequent to this step, binomial probabilities and PWM scores are calculated exactly as they would have been for the full foreground and background data sets as described above. Within a pLogo, fixed residues are depicted by being drawn at full height and on a gray background.

Positioning of the red horizontal pLogo bar. The red horizontal pLogo bar provides a useful reference point when comparing multiple pLogos, and it allows for the quick determination of whether the most statistically significant over- or underrepresented residues in a particular motif position exceed the chosen α value (typically 0.05). To accurately determine placement of the bar, one must correct for the number of statistical tests carried out in the generation of a pLogo. The determination of the number of statistical tests carried out in any given iteration of pLogo generation is equivalent to the number of binomial probabilities calculated and can be determined by the following formula:

$$\text{Number of binomial calculations} = \sum_{i \in R} C_i$$

where R represents the set of positions with nonfixed residues and C_i represents the number of unique characters in the background set at position i. Usually all values of C_i would be equal to 4 for nucleic acid sequences such as DNA or RNA, or 20 for proteins, but this need not be the case, and the formula above is general in nature. Thus, using a conservative Bonferroni correction for an expected α value of 0.05, one would use the following formulas for the placement of the statistical-significance bar on a log scale

$$\alpha' = \frac{0.05}{\sum_{i \in R} C_i}$$

$$\text{Statistical significance bar position} = \pm\log\left(\frac{\alpha'}{1-\alpha'}\right)$$

Note that the placement of the statistical-significance bar is dependent on the number of fixed positions and unique characters in the pLogo.

pLogo tool and website. The current pLogo tool was developed in C++ and provides a simple command-line interface for invoking the tool on a selected foreground and background file. It also accepts input describing motif residues to be fixed as described above. Given these values, the tool calculates the log odds of the binomial probability for each residue at each position in the motif and returns these data to the website for visual pLogo generation. The application was developed using the g++ compiler and GDB for debugging. Binomial-distribution calculations were implemented according to the Incomplete Beta Function method as described previously[18]. All proteomic backgrounds used on the Web tool have been obtained as complete proteomes from the UniProt database[19]. These background proteomes are processed to create a fixed-width background by taking a sliding *n*-mer window across the entire proteome by the pLogo tool. Duplicate *n*-mers are removed in both the foreground and background before pLogo generation.

The back end for the pLogo website was built with the CodeIgniter PHP framework. Probability values for pLogos are calculated by the pLogo tool (mentioned above), and pLogo images are subsequently created from these values by employing the ImageMagick PHP library. The interactive user interface for the website is powered by the jQuery JavaScript library and various open-source jQuery plug-ins. The site runs on an Apache server and utilizes a MySQL database for storing job data. Additional details on the pLogo website will be provided through a subsequent publication and are presently available through a series of instructional videos and help notes on the site.

Sequence visualizations for the iceLogo, WebLogo and Two Sample Logo tools. The iceLogo[9] Java source code was downloaded (http://code.google.com/p/icelogo/) and run locally. The iceLogo program was used in conjunction with the appropriate foreground and background data sets (see main text) to output the iceLogo visualizations shown in **Figures 1** and **2**. Standard sequence logos based on appropriate foreground data sets were generated using the WebLogo[15] tool online (http://weblogo.threeplusone.com/). Two Sample Logos[8] were generated using the Two Sample Logo tool online (http://www.twosamplelogo.org/). Backgrounds uploaded to the Two Sample Logo tool consisted of the same whole proteomic 15-mer backgrounds created by the pLogo tool; however, in the case of the S-centered (**Supplementary Fig. 2**), D-centered (**Supplementary Fig. 4**) and *Drosophila melanogaster* genomic (**Supplementary Fig. 5**) backgrounds, random subsampling was used to prevent "Internal Server Errors" on the Two Sample Logo Web server. Standard default parameters were used to generate WebLogos, iceLogos and Two Sample Logos, which in the case of iceLogos and Two Sample Logos included the visualization of only residues below the significance threshold of 0.05. On occasion it was necessary to alter *y*-axis scales so that no residues would be cropped in the analysis.

Data sources. Src and CaMKII phosphorylation sites (**Figs. 1** and **2** and **Supplementary Tables 1** and **2**) were obtained from the PhosphoSitePlus database[16]. S-nitrosylation sites (**Fig. 1** and **Supplementary Table 1**) were obtained from the literature[20–22]. N-glycosylation sites (**Supplementary Fig. 3** and **Supplementary Table 3**) were obtained from the dbPTM database[23]. Caspase proteolytic cleavage sites (**Supplementary Fig. 4** and **Supplementary Table 4**) were obtained from the literature[24] as well as the Casbah[25] and CutDB[26] databases. P-element insertion sites in *D. melanogaster* (**Supplementary Fig. 5** and **Supplementary Table 5**) were obtained from the literature[27].

18. Press, W.H., Teukolsky, S.A., Vetterling, W.T. & Flannery, B.P. *Numerical Recipes: The Art of Scientific Computing* 3rd edn. (Cambridge University Press, 2007).
19. The UniProt Consortium. *Nucleic Acids Res.* **41**, D43–D47 (2013).
20. Forrester, M.T. *et al. Nat. Biotechnol.* **27**, 557–559 (2009).
21. Doulias, P.T. *et al. Proc. Natl. Acad. Sci. USA* **107**, 16958–16963 (2010).
22. Chen, Y.J., Ku, W.C., Lin, P.Y., Chou, H.C. & Khoo, K.H. *J. Proteome Res.* **9**, 6417–6439 (2010).
23. Lu, C.T. *et al. Nucleic Acids Res.* **41**, D295–D305 (2013).
24. Mahrus, S. *et al. Cell* **134**, 866–876 (2008).
25. Lüthi, A.U. & Martin, S.J. *Cell Death Differ.* **14**, 641–650 (2007).
26. Igarashi, Y. *et al. Nucleic Acids Res.* **35**, D546–D549 (2007).
27. Linheiro, R.S. & Bergman, C.M. *Nucleic Acids Res.* **36**, 6199–6208 (2008).

doi:10.1038/nmeth.2646