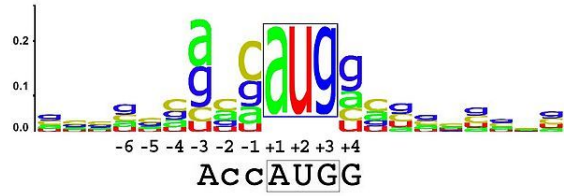
Sequence logo



A sequence logo showing the most conserved bases around the initiation codon from all human mRNAs (Kozak consensus sequence). Note that the initiation codon is not drawn to scale, or the letters AUG would each have a height of 2 bits.

In bioinformatics, a **sequence logo** is a graphical representation of the sequence conservation of nucleotides (in a strand of DNA/RNA) or amino acids (in protein sequences).[1] A sequence logo is created from a collection of aligned sequences and depicts the consensus sequence and diversity of the sequences. Sequence logos are frequently used to depict sequence characteristics such as protein-binding sites in DNA or functional units in proteins.

1 Overview

A sequence logo consists of a stack of letters at each position. The relative sizes of the letters indicates their frequency in the sequences. The total height of the letters depicts the information content of the position, in bits.

2 Logo creation

To create sequence logos, related DNA, RNA or protein sequences, or DNA sequences that have common conserved binding sites, are aligned so that the most conserved parts create good alignments. A sequence logo can then be created from the conserved multiple sequence alignment. The sequence logo will show how well residues are conserved at each position: the higher the number of residues, the higher the letters will be, because the better the conservation is at that position. Different residues at the same position are scaled according to their frequency. The height of the entire stack of residues is the information measured in bits. Sequence logos can be used to represent conserved DNA binding sites, where transcription factors bind.

The information content (y-axis) of position i is given by:

$$R_i = \log_2(20) - (H_i + e_n)$$

$$R_i = 2 - (H_i + e_n)$$

where H_i is the uncertainty (sometimes called the Shannon entropy) of position i

$$H_i = -\sum f_{a,i} \times \log_2 f_{a,i}$$

Here, $f_{a,i}$ is the relative frequency of base or amino acid a at position i , and e_n is the small-sample correction for an alignment of n letters. The height of letter a in column i is given by

$$\text{height} = f_{a,i} \times R_i$$

The approximation for the small-sample correction, e_n , is given by:

$$e_n = \frac{1}{\ln 2} \times \frac{s-1}{2n}$$

where s is 4 for nucleotides, 20 for amino acids, and n is the number of sequences in the alignment.

3 Consensus logo

A **consensus logo** is a simplified variation of a sequence logo that can be embedded in text format. Like a sequence logo, a consensus logo is created from a collection of aligned protein or DNA/RNA sequences and conveys information about the conservation of each position of a sequence motif or sequence alignment[1][2] . However, a consensus logo displays only conservation information, and not explicitly the frequency information of each nucleotide or amino acid at each position. Instead of a stack made of several characters, denoting the relative frequency of each character, the consensus logo depicts the degree of conservation of each position using the height of the consensus character at that position.

3.1 Advantages and drawbacks

The main, and obvious, advantage of consensus logos over sequence logos is their ability to be embedded as

A sequence logo for the LexA-binding motif of several Gram-positive species.

```
(2 bits) |TCGAACANATGTTCGA [IC: 14.748 bits]
```

A consensus logo for the LexA-binding motif of several Gram-positive species.

text in any Rich Text Format supporting editor/viewer and, therefore, in scientific manuscripts. As described above, the consensus logo is a cross between sequence logos and consensus sequences. As a result, compared to a sequence logo, the consensus logo omits information (the relative contribution of each character to the conservation of that position in the motif/alignment). Hence, a sequence logo should be used preferentially whenever possible. That being said, the need to include graphic figures in order to display sequence logos has perpetuated the use of consensus sequences in scientific manuscripts, even though they fail to convey information on both conservation and frequency.[3] Consensus logos represent therefore an improvement over consensus sequences whenever motif/alignment information has to be constrained to text.

4 See also

- Sequence motif
- Position-specific scoring matrix
- DNA binding site

5 References

[1] Schneider TD, Stephens RM (1990). "Sequence Logos: A New Way to Display Consensus Sequences". *Nucleic Acids Res* **18** (20): 6097–6100. doi:10.1093/nar/18.20.6097. PMC 332411. PMID 2172928.

[2] Anzaldi LJ, Muñoz-Fernández D, Erill I. (2012). "BioWord: a sequence manipulation suite for Microsoft Word" (PDF). *BMC Bioinformatics* **13** (124): 124. doi:10.1186/1471-2105-13-124. PMC 3546851. PMID 22676326.

[3] Schneider TD (2002). "Consensus Sequence Zen". *Appl Bioinformatics* **1** (3): 111–119. PMC 1852464. PMID 15130839.

6 External links

- How to read sequence logos.
- Recommendations for Making Sequence Logos.
- Erill, I., "A gentle introduction to information content in transcription factor binding sites", Eprint
- What is (in) a sequence logo?

6.1 Tools for creating sequence logos

- RWebLogo R Code, wrapper for python code (BSD licence)
- WebLogo Python Code Python Code (BSD license, somewhat difficult to use)
- WebLogo 3.0 (Online)
- Seq2Logo (Online app. for peptide alignments feat. pseudo count, sequence weighting and two-sided representation)
- MoRAine (Online application with integrated binding site re-annotation)
- GENIO (Online)
- PWM-based logo (Online application for motif PWM-based models)
- LogoBar (Java application)
- CorreLogo An online server for 3D sequence logos of RNA and DNA alignments
- seqlogo C function to generate DNA sequence logos
- MS-Word AddOn Ribbon that allows generation of consensus logos
- RILogo program and web server for creating logos for two interacting RNAs
- Skylign Online tool for creating logos representing both sequence alignments and profile hidden Markov models

7 Text and image sources, contributors, and licenses

7.1 Text

- **Sequence logo** *Source:* http://en.wikipedia.org/wiki/Sequence_logo?oldid=661897430 *Contributors:* The Anome, Schutz, Peak, WalkinDownThirtyThree, Thorwald, Cacycle, Rjwilmsi, Crystallina, SmackBot, Bluebot, Biehl, Talgalili, ꟼ, TransControl, Alexbateman, Melcombe, Jbening, Addbot, DOI bot, Neodop, Informationtheory, FrescoBot, Gnomehacker, Citation bot 1, Traviswheeler, Cogiati, Max Libbrecht, Tinyraysite, Dexbot, Martin0055 and Anonymous: 14

7.2 Images

- **File:KozakConsensus.jpg** *Source:* http://upload.wikimedia.org/wikipedia/commons/b/bf/KozakConsensus.jpg *License:* CC-BY-SA-3.0 *Contributors:* Transferred from en.wikipedia to Commons. *Original artist:* TransControl at English Wikipedia
- **File:LexA_gram_positive_bacteria_consensus_logo.png** *Source:* http://upload.wikimedia.org/wikipedia/en/1/1e/LexA_gram_positive_bacteria_consensus_logo.png *License:* Cc-by-sa-3.0 *Contributors:*

  Created using the BioWord software (http://compbio.umb.edu/software/bioword/), which is distributed on a GNU GPL license. *Original artist:*

  Gnomehacker (talk) (Uploads)
- **File:LexA_gram_positive_bacteria_sequence_logo.png** *Source:* http://upload.wikimedia.org/wikipedia/commons/8/85/LexA_gram_positive_bacteria_sequence_logo.png *License:* CC BY-SA 3.0 *Contributors:* Source: Created using the Weblogo software (http://weblogo.berkeley.edu/), which is distributed on a MIT Open Source License (http://weblogo.berkeley.edu/LICENSE) *Original artist:* Gnomehacker

7.3 Content license

- Creative Commons Attribution-Share Alike 3.0