H-BloX: visualizing alignment block entropies

Jochen Zuegge,*,† Martin Ebeling,* and Gisbert Schneider*

**F. Hoffmann–La Roche Ltd., Pharmaceuticals Division, Basel, Switzerland*
†Albert-Ludwigs-Universität Freiburg, Institut für Biologie II, Freiburg, Germany

H-BloX is a web-based JavaScript application that allows the calculation and visualization of Shannon information content or relative entropy (Kullback-Leibler 'distance') within sequence alignment blocks. The application was designed for use in both teaching and research. Amino acid, nucleic acid sequences, or any other type of aligned chemical structures may serve as the input. Various interpretations of the meaning of 'entropy' or 'information content' are possible, including treatment as a chemical diversity measure or the degree of feature conservation. For analysis of numerical data by H-BloX, values must be converted to a user-defined character alphabet before computation of entropy or information content. H-BloX was successfully applied to feature identification in Escherichia coli *signal peptides and their cleavage sites. Characteristic known features became visible, e. g., the hydrophobic core region and the well-known '−3,−1' cleavage site pattern. Based on the H-BloX analysis, the hydrophobic core is centered at amino acid residue position 13, counting from the N-terminal end of the protein precursor sequence. This result was obtained by using a built-in feature of H-BloX that enables conversion of amino acid sequences to a different alphabet that is based on hydrophobicity assigments. H-BloX can be accessed online or downloaded as HTML/JavaScript at http://bopwww.biologie.uni-freiburg.de/~bioinfo/HBloX/html/index.html © 2001 by Elsevier Science Inc.*

Keywords: alignment, chemical library, diversity, entropy, information, Shannon

Once a set of isofunctional chemical structures has been compiled, it is highly desirable to understand the underlying function- and structure-determining features. A straightforward approach is the construction of 'alignments' and subsequent analysis of common patterns. The particular alignment procedure chosen depends in part on the type of chemical entities to be investigated. There are several established techniques for amino acid and nucleic acid sequence alignment.[1] We have implemented a computer-based technique, called H-BloX, for graphical display of common patterns present in alignment blocks, which can help reveal relevant features. The Shannon entropy of a sequence alignment block has been proposed as a measure of the randomness of the residue distribution at each aligned position.[1,2] If $P_i(x_k)$ gives the frequency of the symbol x_k from the alphabet $x_1, x_2, x_3, \ldots, x_A$ at the alignment position i, the Shannon entropy at this position is defined by

$$\mathrm{H}_i = -\sum_{k=1}^{A} P_i(x_k)\log_2 P_i(x_k). \tag{1}$$

In this definition, $P_i(x_k)\log_2 P_i(x_k)$ is taken to be zero if $P_i(x_k)=0$. The unit of the Shannon entropy is 'bit' because the base of the logarithm in the formula is two. Choosing the natural logarithm would result in 'nit' units. The Shannon information R_i at the position i in the sequence alignment block is defined as the difference of two entropies:

$$\mathrm{R}_i = \mathrm{H}_{background} - \mathrm{H}_i, \tag{2}$$

where $\mathrm{H}_{background}$ corresponds to the average sequence entropy. $\mathrm{H}_{background}$ is maximized when the symbols of the alphabet are evenly distributed: $P_{background}(x_k)=\frac{1}{A}\ \forall k\epsilon[1,2,3,\ldots,A]$. In this case the formula for $\mathrm{H}_{background}$ simplifies to

$$\mathrm{H}_{background} = \log_2 A, \tag{3}$$

with all calculated values for R_i equal to or larger than zero. Therefore, an even background distribution of all residues is usually assumed, when information theory is applied to sequence alignment blocks. This means that the information content at a given position is high, when the distribution of the symbols is far from random. As a result, calculating the information content for each position in a sequence alignment can be used to spot conserved symbols in the block. However, the naïve assumption of evenly distributed symbols in the background might lead to false interpretations, if the background distribution is highly biased towards certain symbols. Unfortunately, when $\mathrm{H}_{background}$ is calculated using the true background distribution, other problems in the interpretation of R_i

Color Plate for this article is on page 379.

Corresponding author: Gisbert Schneider, F. Hoffmann–La Roche Ltd., Pharmaceuticals Division, CH-4070 Basel, Switzerland. Tel.: +41 61 68 70696, fax: +41 61 68 87408.

E-mail address: gisbert.schneider@roche.com

Journal of Molecular Graphics and Modelling 19, 304–306, 2001

655 Avenue of the Americas, New York, NY 10010

1093-3263/01/$–see front matter
PII S1093-3263(00)00074-7

might occur. It could happen, for example, that R_i is calculated to be zero, although the frequency of each symbol differs in $H_{background}$ and H_i, simply because both distributions are equally far from a totally random distribution. This is because R_i only tells something about whole distributions of symbols, without comparing the frequency of each specific symbol directly. As a possible solution to this problem, one can calculate the relative entropy $H_i(P_i \| P_{background})$, also known as Kullback-Leibler 'distance,' defined by

$$H_i(P_i \| P_{background}) = \sum_{k=1}^{A} P_i \log_2 \frac{P_i(x_k)}{P_{background}(x_k)}. \qquad (4)$$

The relative entropy equals the Shannon information for an even background distribution, but differs otherwise. $H_i(P_i \| P_{background})$ is always equal to or greater than zero. It vanishes only, if every single symbol has the same frequency in both the background distribution and within a sequence block position. Contrary to the Shannon entropy, relative entropy is not a 'state function',[3] and although it is often useful to think of relative entropy as a distance between two probability distributions, it is not symmetric and is not a correct mathematical distance measure.[1] For potential applications of Equation 4, see, for example, the textbooks of Durbin et al.[1] and Baldi and Brunak.[4] The H-BloX web-application allows the user to calculate the Shannon information or the relative entropy at each position of an ungapped alignment block. The creation of the block itself is not a part of the program. However, there are other freely accessible web-applications that may accomplish this task.[5,6] H-BloX can interpret input alignment blocks in the FASTA and BLOCKS format as well as raw data. Other common sequence formats may be converted to the FASTA format using, e.g., the 'ReadSeq' web application.[7] The results of the H-BloX calculations are presented in a color-coded output table. There are alphabet presets for proteins, DNA and RNA. Additionally, it is possible to define a custom alphabet or let the program find the used alphabet in the block by itself. A custom-tailored background distribution of the alphabet symbols may be defined, if desired. There are several background distributions of amino acids predefined for different taxa (computed from the SWISSPROT database, release 38). A special feature of H-BloX is the possibility to translate the alphabet (and the sequence alignment) to a different, user-defined alphabet prior to entropy calculation. This can be very useful in revealing underlying patterns, which are invisible considering the original alphabet. For example, H-BloX offers the possibility to distinguish between 'hydrophobic' (A, C, F, G, I, L, M, T, V, W) and 'other' (D, E, H, K, N, P, Q, R, S, Y) amino acid residues (grouping based on the hydrophobicity scale according to Engelman et al.).[8] This leads to a new alphabet consisting only of the two symbols 'H' (hydrophobic) and 'O' (other). It must be kept in mind that the background frequencies of the new symbols are dependent on both the original distribution and the translation.

To demonstrate the usefulness of this option, two H-BloX applications are shown as an example in Color Plate 1. First, it is known that N-terminal secretion signals (signal peptides) usually contain a hydrophobic core region followed by a cleavage site region; with positions -3 and -1 relative to the signal peptidase cleavage site occupied by small and neutral amino acids.[9] The information content of the cleavage site region has been investigated recently.[10] In Color Plate 1a the Shannon information (Equation 2) is shown for the Nielsen collection of *Escherichia coli* (*E. coli*) signal peptidase cleavage site sequences, assuming equal background frequencies of the 20 standard residues. The H-BloX output clearly shows the expected pattern, including a section of hydrophobic core and the '-3, -1' box.[9,10] Interestingly, the pattern becomes much clearer, when the sequences are translated to the hydrophobic/other alphabet (Color Plate 1b). Especially at relative position -6 (position 10 in Color Plate 1a,b) the output of the H-BloX program differs depending on the particular alphabet selected. The information content is significantly above zero with the standard amino acid single letter code as the alphabet, indicating that some residues are predominant at this position (Color Plate 1a). Position –6 marks the borderline between the hydrophobic core and the cleavage site region (Color Plate 1b). This observation is substantiated by several other investigations.[9,11] As a second example, taking the N-terminal parts of potentially secreted proteins from *E.coli* as input for calculation of the Shannon information content (Equation 2; equal background frequencies assumed), demonstrates that only the conserved N-terminal methionines lead to high bit values (Color Plate 1c). If, however, the sequences are converted to the hydrophobic/other alphabet, the hydrophobic core of signal peptides becomes visible (Color Plate 1d). Based on this analysis its average mid-point is located at sequence position 13 (counting from the N-terminal end), as indicated by the high information content value. An additional feature of H-BloX is the calculation and graphical display of a histogram giving the difference between the expected and the actual residue frequencies in the alignment block (not shown). H-BloX can either be used directly on the web page, or downloaded as a HTML/JavaScript for local use (Netscape browser version 4.x or higher recommended). It was especially designed for bioinformatics classes, and as an easy-to-use research tool. Full documentation is supplied within the source code, enabling easy modification and extension.

CONCLUSIONS

Calculation of alignment block information content can be of considerable value for finding underlying 'hidden' sequence patterns. H-BloX provides a freely accessible and easy-to-use interface for this purpose. Its usefulness was demonstrated taking the analysis of signal peptide features as an example. Ungapped alignment blocks form the basis of the BLOCKS and PRINTS protein motif databases, which can be browsed over the web.[5,6] From these databases alignment blocks corresponding to known, conserved protein motifs can be retrieved and used as input for H-BloX. The BLOCKS site also allows for the construction of new alignment blocks from user-defined sequence data. MEME is a software tool set up for the identification of conserved motifs in unaligned input sequences.[12] MEME's output - primarily a profile to be used in database searches - also contains ungapped alignment blocks of the identified motifs. For manual creation and alignment editing the programs CINEMA and Jalview are available on the web.[13,14] Although these software packages provide a sophisticated range of viewing options they are not capable of showing the information content of an ungapped sequence block or translate the alphabet of the aligned sequences to another. H-BloX meets this need and is thought to complement similar tools. The H-BloX analysis is not restricted to protein, DNA or RNA sequences. It may be applied to arbitrary chemical libraries as well, provided a sensible alignment of structures can be accomplished. As long as the molecular structure can be described by a limited set of building blocks, which is often possible for combinatorial libraries, a corresponding sequence-representation may be used as the input data for H-BloX. In this context the term 'entropy' can be interpreted as a measure of 'library diversity'.[15]

ACKNOWLEDGMENTS

We thank Uwe Hobohm and Karin Gierke for helpful discussions and comments on the manuscript.

REFERENCES

1 Durbin, R., Eddy, S., Krogh, A., and Mitchinson, G. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids.* Cambridge University Press, Cambridge, UK, 1998

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

3 Schneider, T.D. Measuring molecular information. *J. Theor. Biol.* 1999, **201**, 87–92

4 Baldi, P., and Brunak, S. *Bioinformatics - The Machine Learning Approach.* MIT Press, Cambridge, USA, 1998

5 Henikoff, S., Henikoff, J.G., Alford W.J., and Pietrokovski, S. Automated construction and graphical presentation of protein blocks from unaligned sequences. *Gene-COMBIS, Gene* 1995, **163**, 17–26 (http://blocks.fhcrc.org/blocks/blockmkr/make−blocks.html)

6 Attwood, T.K., Croning, M.D., Flo, D.R., Lewis, A.P., Mabey, J.E., Scordis, P., and Selley, J.N. PRINTS-S: the database formerly known as PRINTS. *Nucleic Acids Res.* 2000, **28**, 225–227 (http://www.bioinf.man.ac.uk/dbbrowser/PRINTS/)

7 Worley, K.C. Convert Sequence Formats using ReadSeq (http://dot.imgen.bcm.tmc.edu:9331/seq-util/readseq.html)

8 Engelman, D.M., Steitz, T.A., and Goldman, A. Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. *Annu. Rev. Biophys. Biophys. Chem.* 1986, **15**, 321–353

9 von Heijne, G. Signal sequences. The limits of variation. *J. Mol. Biol.* 1985, **184**, 99–105

10 Nielsen, H., Engelbrecht, J., Brunak, S., and von Heijne, G. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng.* 1997, **10**, 1–6 (ftp://virus.cbs.dtu.dk/pub/signalp/)

11 Schneider, G., Röhlk, S., and Wrede, P. Analysis of cleavage-site patterns in protein precursor sequences with a perceptron-type neural network. *Biochem. Biophys. Res. Commun.* 1993, **194**, 951–959

12 Grundy, W.N., Bailey, T.L., and Elkan, C.P. ParaMEME: A Parallel Implementation and a Web Interface for a DNA and Protein Motif Discovery Tool, *Computer Applications in the Biological Sciences (CABIOS)* 1996, **12**, 303–310 (http://meme.sdsc.edu/meme/website)

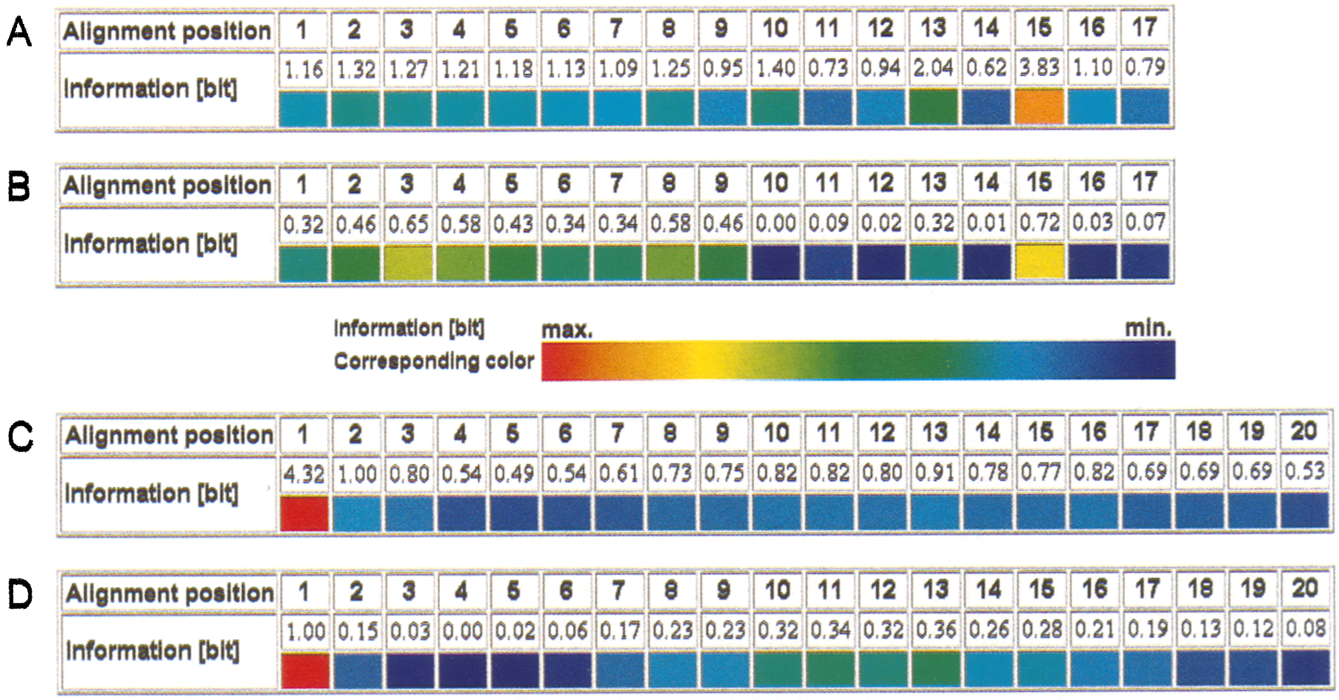
13 Parry-Smith, D.J., Payne, A.W., Michie, A.D., and Attwood, T.K. CINEMA—a novel colour interactive editor for multiple alignments. *Gene* 1998, **221**, 57–63 (http://bioinformatics.weizmann.ac.il/CINEMA/)

14 Clamp, M. Jalview - a java multiple alignment editor (http://jura.ebi.ac.uk:6543/jalview/)

15. Böhm, HJ and Schneider, G (Eds.). *Virtual screening for bioactive molecules.* Wiley/VCH, Weinheim, 2000

Jochen Zuegge, Martin Ebeling, and Gisbert Schneider

H-BloX: visualizing alignment block entropies K-Ras4B binding specificity to protein farnesyl-transferase revealed by 2 Å resolution ternary complex structures. *Structure*


A

Alignment position	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Information [bit]	1.16	1.32	1.27	1.21	1.18	1.13	1.09	1.25	0.95	1.40	0.73	0.94	2.04	0.62	3.83	1.10	0.79

B

Alignment position	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Information [bit]	0.32	0.46	0.65	0.58	0.43	0.34	0.34	0.58	0.46	0.00	0.09	0.02	0.32	0.01	0.72	0.03	0.07

Information [bit] max. min.
Corresponding color

C

Alignment position	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Information [bit]	4.32	1.00	0.80	0.54	0.49	0.54	0.61	0.73	0.75	0.82	0.82	0.80	0.91	0.78	0.77	0.82	0.69	0.69	0.69	0.53

D

Alignment position	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Information [bit]	1.00	0.15	0.03	0.00	0.02	0.06	0.17	0.23	0.23	0.32	0.34	0.32	0.36	0.26	0.28	0.21	0.19	0.13	0.12	0.08

Color Plate 1. Examples of color-coded patterns of Shannon information values produced by H-BloX. (A) analysis of 105 aligned *E.coli* leader peptidase substrate sequences encompassing positions −15 to +2 relative to the cleavage site. (B) same data as in (A), where the residue alphabet was translated to 'Hydrophobic' and 'Other' prior to calculation of the information content. (C) information pattern derived from N-terminal parts (20 residues) of 487 potentially secreted protein precursors from *E.coli*; (D) same data as in (C) with the hydrophobic/other residue translation.