BIOINFORMATICS ORIGINAL PAPER

Vol. 31 no. 3 2015, pages 324–331
doi:10.1093/bioinformatics/btu634

Sequence analysis

Advance Access publication October 6, 2014

Log-odds sequence logos

Yi-Kuo Yu[1], John A. Capra[2,3], Aleksandar Stojmirović[1,†], David Landsman[1] and Stephen F. Altschul[1,*]

[1]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, [2]Center for Human Genetics Research and [3]Department of Biomedical Informatics, Vanderbilt University, Nashville, TN 37232, USA

Associate Editor: Alfonso Valencia

ABSTRACT

Motivation: DNA and protein patterns are usefully represented by sequence logos. However, the methods for logo generation in common use lack a proper statistical basis, and are non-optimal for recognizing functionally relevant alignment columns.

Results: We redefine the information at a logo position as a per-observation multiple alignment log-odds score. Such scores are positive or negative, depending on whether a column's observations are better explained as arising from relatedness or chance. Within this framework, we propose distinct normalized maximum likelihood and Bayesian measures of column information. We illustrate these measures on High Mobility Group B (HMGB) box proteins and a dataset of enzyme alignments. Particularly in the context of protein alignments, our measures improve the discrimination of biologically relevant positions.

Availability and implementation: Our new measures are implemented in an open-source Web-based logo generation program, which is available at http://www.ncbi.nlm.nih.gov/CBBresearch/Yu/logoddslogo/index.html. A stand-alone version of the program is also available from this site.

Contact: altschul@ncbi.nlm.nih.gov

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on March 19, 2014; revised on August 26, 2014; accepted on September 18, 2014

1 INTRODUCTION

Patterns or motifs that are shared among multiple DNA or protein sequences often correlate with important biological functions. A widely used method for representing and studying such patterns is provided by 'sequence logos' (Crooks *et al.*, 2004; Schneider and Stephens, 1990). A sequence logo represents concisely two important features of a multiple alignment—the 'information content' at each alignment position and the frequencies of the nucleotides or amino acids observed at that position. Here, we suggest modifications to the way in which the information content is usually measured. These modifications yield a direct connection between the information for an alignment position and a multiple alignment log-odds score (Altschul *et al.*, 2010). We have implemented a program that uses the modified information measures, and illustrate its application to protein alignments.

*To whom correspondence should be addressed.
†Present address: Janssen Research & Development, LLC, Spring House, PA 19477, USA.

2 METHODS

2.1 Background

Sequence logos were introduced by Schneider and Stephens (1990), and a popular Web-based logo construction program was made available by Crooks *et al.* (2004). A variety of alternative methods for calculating and presenting logos have since been described (Colaert *et al.*, 2009; O'Shea *et al.*, 2013; Schuster-Böckler *et al.*, 2004; Vacic *et al.*, 2006; Workman *et al.*, 2005). A logo is derived from a multiple alignment, and represents each position of the alignment by a stack of letters, each corresponding to a nucleotide or an amino acid. The height of each letter in the stack is proportional to the observed frequency of the letter at that position, and the aggregate height of the stack corresponds to the 'information content' of the position.

To formalize, assume our alphabet has size L (4 for DNA, and 20 for proteins), indexed by the first L positive integers, and that the background frequencies with which these letters appear by chance is given by the fixed vector $\vec{p}$. We represent by $\vec{u}$ the special uniform case in which all $p_j = 1/L$. Assume at a given alignment position we have N observations, given by the vector $\vec{x}$, with $x_i \in \{1, \ldots, L\}$. Let the counts of the various letters within $\vec{x}$ be given by the vector $\vec{n}$, and let us define the observed frequencies as $\vec{q} \equiv \vec{n}/N$.

From the information theory (Cover and Thomas, 1991), the entropy (in bits) of a frequency distribution $\vec{f}$ is defined as $H(\vec{f}) \equiv -\sum_{j=1}^{L} f_j \log_2 f_j$. Schneider and Stephens (1990) use entropy to define the 'information content' of a given position as the difference between the maximum possible entropy and that observed at the position as follows:

$$R(\vec{x}) \equiv H(\vec{u}) - \left[H(\vec{q}) + e(N)\right] = \log_2 L - H(\vec{q}) - e(N). \quad (1)$$

Here, $e(N)$ is a correction for small sample size (Schneider *et al.*, 1986) (see Appendix, Supplementary Material), which approaches 0 for large N. To calculate $e(N)$ using the approach of (Schneider *et al.*, 1986), but without assuming $\vec{p} = \vec{u}$, requires summing $C_N^{N+L-1} = \mathcal{O}(N^{L-1})$ terms. For proteins, this becomes impractical well before N reaches 30, so the term $e(N)$ has generally been ignored (Crooks *et al.*, 2004), giving rise to the measure $U(\vec{x})$:

$$U(\vec{x}) \equiv H(\vec{u}) - H(\vec{q}) = \log_2 L - H(\vec{q}). \quad (2)$$

Because expectation and summation commute, however, one may calculate $e(N)$ exactly for arbitrary $\vec{p}$ in time $\mathcal{O}(LN)$, and its standard deviation in time $\mathcal{O}(L^2N^2)$; all factors of L may be dropped when $\vec{p} = \vec{u}$. This does not appear to have been observed before. Accordingly, while we study both $R(\vec{x})$ and $U(\vec{x})$ in this article, only $U(\vec{x})$ has previously been in widespread use for proteins.

Our proposal is simply to replace $R(\vec{x})$ or $U(\vec{x})$ with one of the two alternative measures, $A(\vec{x})$ or $B(\vec{x})$, as defined and discussed below. Although the proposed change may appear minor, in the protein alignment context it can have important consequences for the appearance, interpretation and use of sequence logos.

The WebLogo program (Crooks *et al.*, 2004) implements a number of modifications to the original theory (Schneider and Stephens, 1990). These modifications are not described in detail either in Crooks *et al.* (2004) or in the program's user manual, but examination of the code reveals they are different in substance and spirit from those we advance here. Thus, for brevity, accuracy and clarity, we will compare our proposed measures of column information only with the original measure $R(x)$ of Schneider and Stephens (1990) and its uncorrected form $U(\vec{x})$. The measure implemented by WebLogo is the closest in form and output to $U(\vec{x})$.

2.2 The log-odds perspective

The average per-observation information available for distinguishing a stretch of symbols that follows a known probability distribution $\vec{q}$ from a stream of symbols that follows a background distribution $\vec{p}$ is generally understood to be the distributions' relative entropy (Cover and Thomas, 1991):

$$D(\vec{q}||\vec{p}) \equiv \sum_{j=1}^{L} q_j \log_2 \frac{q_j}{p_j}. \tag{3}$$

By continuity, one may set $0\log_2 0 \equiv 0$, and thus render D well-defined for $\vec{q}$ in which some components are zero. The asymmetry of D in $\vec{q}$ and $\vec{p}$ can be understood by considering the relative ease of recognizing a stretch of a dozen fair coin flips within a sea of million that are 99.9% biased toward heads, compared with the converse.

To recognize a DNA motif within a long stretch of DNA is a similar problem, except that the frequency of nucleotides within the motif differs from the background frequencies in a position-dependent manner. In the infinite N limit, a motif's frequencies at an individual position may be taken as a known, fixed vector $\vec{q}$, so the average information available per observation should be given by Equation (3).

In the case of uniform background frequencies,

$$\begin{aligned} D(\vec{q}||\vec{u}) &= \sum_{j=1}^{L} q_j \log_2 \frac{q_j}{u_j} = \sum_j q_j \log_2 q_j - \sum_j q_j \log_2 u_j \\ &= \log_2 L - H(\vec{q}), \end{aligned} \tag{4}$$

so that the relative-entropy measure of column information is numerically equal to Schneider and Stephens' measure. However, the two definitions differ substantially when $\vec{p} \neq \vec{u}$, which is the case for many organisms of interest. Most obviously, for a column dominated by a particular letter a, so that $q_a \approx 1$, the R of Equation (1) approaches $\log_2 L$ bits, whereas the D of Equation (3) approaches $\log_2 1/p_a$ bits. In other words, relative entropy yields higher information content for columns dominated by a letter with lower background frequency. Schneider *et al.* (1986) suggested relative entropy as an alternative to **1** for measuring column information, but Schneider and Stephens (1990) did not pursue this suggestion further. Relative entropy has also been advocated by Lawrence *et al.* (1993), Schuster-Böckler *et al.* (2004), Stormo (1998) and Workman *et al.* (2005). Crooks *et al.* (2004) have made the same suggestion, but contrary to their statement, replacing $\vec{u}$ by $\vec{p}$ in Equation (1) does not render the two measures equivalent.

Before turning to the small N case, it is useful to observe that, given a $\vec{q}$ exactly proportional to $\vec{n}$, relative entropy can be understood as the per-observation log-odds ratio of $Q(\vec{x})$ to $P(\vec{x})$, where $Q(\vec{x})$ and $P(\vec{x})$ are the likelihoods of generating $\vec{x}$ given the two alternative multinomials $\vec{q}$ and $\vec{p}$. In other words,

$$\frac{1}{N}\log_2 \frac{Q(\vec{x})}{P(\vec{x})} = \frac{1}{N}\log_2 \frac{\prod_j q_j^{n_j}}{\prod_j p_j^{n_j}} = \frac{1}{N}\sum_j n_j \log_2 \frac{q_j}{p_j} = D(\vec{q}||\vec{p}). \tag{5}$$

By requiring, for small N, that Q remain a probability distribution over $\vec{x}$, the log-odds perspective provides the key to our proposed measures of column information.

The special status of log-odds scores has long been recognized in the local alignment context (Altschul, 1991; Karlin and Altschul, 1990). There, the formalism is used explicitly to construct all popular pairwise substitution matrices (Henikoff and Henikoff, 1992; Schwartz and Dayhoff, 1978). A negative log-odds score simply means that the aligned letters are explained better by a model of chance than a model of relatedness. There are many advantages to extending log-odds scores to multiple alignments, as discussed at length in Altschul *et al.* (2010). They may, for example, be used in a principled way to distinguish related from unrelated alignment regions and to include or exclude sequences from a multiple alignment (Altschul *et al.*, 2010). In this article, we show empirically that even within alignments of related protein sequence regions, the relative importance of alignment positions, which logos are designed to represent, is better captured using log-odds scores than other measures of column information in common use.

2.3 The normalized maximum-likelihood-based column information measure

For finite N, although the background $\vec{p}$ remains fixed, $\vec{q}$ is taken to be the maximum-likelihood multinomial $\vec{n}/N$ implied by the observation $\vec{x}$, and thus varies from one $\vec{x}$ to another. For a fixed multinomial, the likelihoods for all possible $\vec{x}$ sum to 1, and may therefore be considered probabilities, but this is no longer the case if $\vec{q}$ varies with $\vec{x}$. Instead, this sum is given by $Z \equiv \sum_{\vec{x}} \prod_j (n_j/N)^{n_j}$. In minimum description length theory, $\log_2 Z$ is called the complexity of the multinomial model, and the measure $A(\vec{x})$ described below can be seen as derived from the normalized maximum-likelihood description length of $\vec{x}$ (Grünwald, 2007).

To convert the likelihoods for the $\vec{x}$ implied by $\vec{q}(\vec{x})$ into probabilities, they must be divided by Z, so that we must write $Q(\vec{x}) = (\prod_j q_j^{n_j})/Z$. Thus, the log-odds perspective of Equation (5) leads us to define our per-observation log-odds measure of column information as follows:

$$A(\vec{x}) \equiv \left(\sum_j q_j \log_2 \frac{q_j}{p_j} \right) - c(N), \tag{6}$$

where $c(N) \equiv \frac{1}{N}\log_2 \sum_{\vec{x}} \prod_j q_j^{n_j}$ approaches 0 for large N and plays an analogous but logically and numerically distinct role to the term $e(N)$ in Equation (1). We give values of $c(N)$ for small N in Supplementary Tables S1 and S2, and derive in the Appendix (see Supplementary Material) an asymptotic formula for $c(N)$. Interestingly, for large N and $\vec{p} = \vec{u}$, $e(N) \approx c(N)/\ln N$ (Altschul *et al.*, 2009; Schneider *et al.*, 1986). The $R(\vec{x})$ that results from Equation (1) has expected value 0 for columns of random letters, whereas in this context, $A(\vec{x})$ has negative expected value, because of its derivation as a per-observation log-odds score (Altschul *et al.*, 2010; Karlin and Altschul, 1990).

2.4 The BILD-score-based column information measure

The approach above calculates $Q(\vec{x})$ by inferring a maximum-likelihood multinomial $\vec{q}(\vec{x})$ for each $\vec{x}$, and normalizing the resulting likelihoods. An alternative Bayesian approach is to define a prior probability density, over the space of all L-letter multinomials, for the $\vec{q}$ associated with a motif column. $Q(\vec{x})$ may then be taken as the probability of observing $\vec{x}$ implied by this prior. Because no parameters are fitted, no correction term analogous to $c(N)$ above is needed.

This approach is mathematically tractable if the prior is chosen as a Dirichlet mixture (Altschul *et al.*, 2010; Brown *et al.*, 1993; Sjölander

et al., 1996). Such a mixture, the superposition of M Dirichlet components, is specified by its mixture parameters $\vec{m}$ (M positive numbers that sum to 1), and its Dirichlet parameters $\vec{\alpha}_{i,j}$ (positive numbers indexed by $i \in \{1, \ldots, M\}$ and $j \in \{1, \ldots, L\}$). It is convenient to define 'concentration parameters' $\alpha_i \equiv \sum_j \alpha_{i,j}$. The resulting measure of column information is given by the following:

$$B(\vec{x}) \equiv \frac{1}{N} \log_2 \frac{Q(\vec{x})}{\prod_j p_j^{n_j}} \tag{7}$$

where

$$Q(\vec{x}) = \sum_{i=1}^{M} m_i \frac{\Gamma(\alpha_i)}{\Gamma(\alpha_i + N)} \prod_{j=1}^{L} \frac{\Gamma(\alpha_{i,j} + n_j)}{\Gamma(\alpha_{i,j})}. \tag{8}$$

This is the per-observation 'Bayesian Integral Log-odds' or 'BILD' score defined in Altschul *et al.* (2010).

A more indirect Bayesian approach, which we have not implemented, is first to infer, from the prior and the observations, a posterior multinomial $\hat{\vec{q}}$ (Altschul *et al.*, 2010; Brown *et al.*, 1993; Sjölander *et al.*, 1996) and then to calculate a column-information score based on $\hat{\vec{q}}$. This may be done using Equation (6), but with its second q_j replaced by $\hat{q}_j$, and with q_j replaced by $\hat{q}_j$ in the formula for $c(N)$. The correction term $c(N)$ now depends on the priors used, so the values given in the Supplementary Tables S1 and S2 no longer apply, nor do the constants derived in the Appendix for the asymptotic formula for $c(N)$. Minus its log-odds correction for sample size, this alternative Bayesian approach has strong similarities to the HMM measure proposed in Schuster-Böckler *et al.* (2004). For proteins, both use Dirichlet mixtures to take account of the prior knowledge concerning amino acid relationships, as do BILD scores. However, as we will see below, properly correcting scores for small sample size (a distinct issue) is important if one wishes to compare the scores of columns containing differing numbers of observations.

2.5 Relationship of measures *A* and *B*

A fairly deep connection exists between measures A and B. To a first approximation, the two measures are equal when the Bayesian prior required for B is taken as a single Dirichlet with all $\alpha_j = 1/2$ (Grünwald, 2007). This is known as the Jeffreys prior, and is uninformative in a truer sense than the uniform prior, which has all $\alpha_j = 1$ (Grünwald, 2007; Jeffreys, 1946). Thus, measure A may be seen, in essence, as a special case of measure B, in which no prior knowledge concerning relationships among the letters is available.

If there is substantial prior knowledge about which letters are likely to co-occur in columns representing true relationships, then by capturing this knowledge through its priors, the BILD-score-based measure of Equations (7) and (8) should, for related columns, usually exceed the normalized maximum-likelihood-based measure of Equation (6). Because such prior knowledge is available for proteins, we prefer $B(\vec{x})$ to $A(\vec{x})$ in this context. It is less clear that B will have any substantial advantage to A for DNA.

2.6 Logo boundaries

Basing measures of column information on log-odds multiple alignment scores provides an automatic means for determining boundaries for a logo (Altschul *et al.*, 2010). Because log-odds scores represent the information for distinguishing true from chance similarities, it is appropriate to seek logos with maximum implied log-odds score (Altschul *et al.*, 2010; Karlin and Altschul, 1990). This is achieved by choosing a contiguous set of columns that maximizes $\sum_k N_k A(\vec{x}_k)$ or $\sum_k N_k B(\vec{x}_k)$, where N_k is the number of observations in column k, and $\vec{x}_k$ is the associated observation vector. This optimization may be achieved using the algorithm of Smith and Waterman (1981), simplified to the one-dimensional case.

2.7 Sequence weights

The mathematics underlying multiple alignment log-odds scores (Altschul *et al.*, 2010) makes the implicit assumption that the letters observed in a column are independent draws from the multinomial representing that column. This is a reasonable assumption for most DNA multiple alignments, where the constraints on sequence are primarily physical (as in the alignment of instances of a transcription factor's binding sites across a genome) rather than historical. However, many or most protein multiple alignments contain some sequences that are similar primarily due to recent ancestry rather than to physical constraints. This problem may be mitigated through sequence weighting (e.g Altschul *et al.*, 1989; Henikoff and Henikoff, 1994; Sunyaev *et al.*, 1999). Such weighting should account not only for the relative frequencies of the residues observed in each column but also for the effective number of independent observations as well (Altschul *et al.*, 1997, 2009; Brown *et al.*, 2007; Sunyaev *et al.*, 1999). Here, we use the weighting method of Sunyaev *et al.* (1999) as modified in Altschul *et al.* (2009). The method yields aggregate numbers of independent observations that vary by column, even for alignments without gaps; this result is consistent with the idea that evolution proceeds at different rates at different protein positions.

2.8 Program

We have modified the open-source WebLogo code (Crooks *et al.*, 2004) to implement the definitions of column information presented above. The new program, LogOddsLogo, is available at http://www.ncbi.nlm.nih.gov/CBBresearch/Yu/logoddslogo/index.html

A stand-alone version of the program is also available from this site. The program's most important features are as follows:

(1) The user may select either $A(\vec{x})$ of Equation (6) or $B(\vec{x})$ of Equations (7) and (8) to calculate the information content of sites. In either case, this information may be understood as a per-observation log-odds score. The measures $R(\vec{x})$ and $U(\vec{x})$ may also be specified. Null or gap characters are ignored, and not counted toward the number of observations in a column.

(2) Because log-odds scores measure the relative evidence for models of relatedness and chance, information may be positive or negative, and is thus represented in the logo by positive or negative departures from the axis.

(3) For proteins, if amino acid counts are calculated from a multiple alignment, as opposed to being supplied explicitly, the user has the option of automatically calculating weighted observations (Altschul *et al.*, 2009; Sunyaev *et al.*, 1999) for each residue.

To apply measure B, one needs to specify a Dirichlet mixture prior. For DNA logos, we take this prior to be a single Dirichlet distribution, with parameters $\alpha_j = \alpha p_j$. The user may specify the concentration parameter α, but we set $\alpha = 1$ by default, as recommended in Nishida *et al.* (2009). As in WebLogo (Crooks *et al.*, 2004), the background frequencies p_j are uniform by default, but the user may select those characteristic of various model organisms, or specify an arbitrary cytosine-guanine (CG) percentage. To apply measure B to proteins, the user may select among several Dirichlet mixtures, and amino acid background frequencies are then inferred from the mixture chosen. For other measures, the background frequencies of Robinson and Robinson (1991) are used. The stand-alone version of the program also allows the user to specify arbitrary background amino acid frequencies.

3 RESULTS

In the DNA context, it is appropriate to posit ignorance and use Jeffreys priors, essentially rendering measures A and B equivalent. Furthermore, although these measures are formally

distinct from the entropy difference measure (Schneider and Stephens, 1990), for uniform background frequencies, they differ only by an N-dependent constant, which approaches 0 as N grows. Measures A and B do differ substantially from entropy difference in the important case of non-uniform background frequencies, but for large datasets, they are still similar to the known relative-entropy measure. Thus, it is primarily in the protein context that a user will encounter noticeable differences between the measures we propose and other measures in common use. Here, to apply measure B to proteins, we use the Dirichet mixture 'dist-ncbi134', described in Nguyen *et al*. (2013). For the sake of consistent comparison, we use the background frequencies implied by this mixture throughout.

When used to analyze proteins, logos are usually applied to multiple alignments in which all the sequence segments included are, in fact, related. The primary purpose is to visualize those positions that are the most conserved, and thus presumably the most relevant to protein structure and function. To illustrate the behavior of our various measures, we first apply them to a particular HMGB box alignment in which the most important positions may be independently determined by structural features and by conservation within a broader encompassing family. Second, we consider a set of alignments of enzyme families, in which the catalytic sites stand as proxies for the most important positions.

3.1 HMGB box proteins

We consider a subset of the HMGB box family of DNA binding proteins (http://www.ncbi.nlm.nih.gov/Structure/cdd/cddsrv.cgi?hslf=1&uid=cd01388&#seqhrch) prepared from the Conserved Domain Database (CDD) at NCBI (Marchler-Bauer *et al*., 2013).

We downloaded 24 aligned domains of the HMGB box domain from the SOX-TCF_HMG-box subfamily for use in this analysis (Supplementary Table S3). These subsequences comprise the ~69–72 amino acid long DNA-binding domain, and were aligned with reference to the *Saccharomyces cerevisiae* protein NHP6A (Masse *et al*., 2002) with known structure (PDB: 1J5N http://www.ncbi.nlm.nih.gov/Structure/mmdb/mmdbsrv.cgi?uid=49732).

The HMGB box proteins form a large family with a wide range of DNA-binding functions and specificities. A subgroup within the family consists mostly of DNA-binding proteins that bind to specific target sequences, and the structures of several of the domains have been determined. For example, the structure of NHP6A (Masse *et al*., 2002) is similar to that of other HMGB box protein domains such as TCF/LEF-1 and SRY (Stros, 2010). The compact domain contains a conserved sequence motif described in Landsman and Bustin (1993). These sequences fold into an L-shaped structure with three alpha helices and an unstructured N-terminal peptide. Using various considerations described in the caption of Figure 1, and illustrated in the Supplementary Figure S1, a set of 18 positions, numbered 5–10, 12, 17, 25, 29, 37, 40, 43, 47, 48, 51, 62 and 69 within our alignment, can be identified as particularly important. Below, we will consider these 18 positions to be either true-positives or false-negatives, and all other positions either false-positives or true-negatives.

The four panels of Figure 1 show sequence logos for our HMGB box alignment that were constructed, respectively, using the information measures U, R, A and B. We weighted sequences using the method of Sunyaev *et al*. (1999) as modified in Altschul *et al*. (2009). All four measures report scores in bits, and are shown using a common scale, but only measures A and B, which have a similar log-odds interpretation, are directly comparable. For each measure, we selected a threshold score that yields seven false-positives, and the corresponding logo positions are colored blue. (Despite their designation here as 'false-positives', these strongly conserved positions may well have important biological functions that have not yet been described.) Using these thresholds, we have colored true-positive positions red and false-negative positions orange. The measures U, R, A and B yield, respectively, 13, 13, 15 and 18 true-positives. The use of seven false-positives here is arbitrary. Different choices would yield different threshholds for separating true-positives from false-negatives, but broadly similar qualitative results.

Comparing the panels of Figure 1 for measures U and R, it is evident that by correcting for the noise from small sample sizes, measure R renders the most important positions more visually prominent. However, for an alignment in which all columns contain the same number N of independent observations, measures U and R differ only by the constant correction term $e(N)$, and thus, with corresponding thresholds, will always yield the same false- and true-positives. Because the weighting system we use (Altschul *et al*., 2009; Sunyaev *et al*., 1999) yields estimates for N that vary by column, measures U and R yield true-positives and false-negatives that are not completely congruent, although nevertheless similar. As we will see below, these measures can differ greatly when applied to sets of alignments containing widely varying numbers of sequences.

Measure A differs from R in two major respects. First, its correction term $c(N)$ is generally greater than $e(N)$. This results in a more complete visual suppression of noise from non-conserved positions. Second, A attributes relatively greater information to positions populated primarily by rare amino acids. Comparing the panels of Figure 1 for measures R and A, this results in the elevation to true-positives of positions 10, 12, 62 and 69, whose most frequent amino acids are, respectively, the relatively rare M, F, H and Y. At the same time, it results in the demotion to false-negatives of positions 37 and 47, whose most frequent amino acids are the relatively common G and E, respectively.

Among our measures, B is the only one to exploit prior amino acid relationship information. For alignments of related sequences, B thus yields greater reported information than A at almost all positions. This information increase will not necessarily be greater for important positions, but in our example it elevates to true-positive positions 37, 47 and 48. This stems, respectively, from the recognized relationships among the small G and A, the polar and negatively biased E, Q and D and the polar and positively biased K, Q and R (Brown *et al*., 1993; Nguyen *et al*., 2013).

One may not conclude from this single example that measure B will on average outperform A at recognizing important alignment positions, or that A will outperform U and R. However, the example illustrates the qualitative features that differentiate these measures and affect their performance.

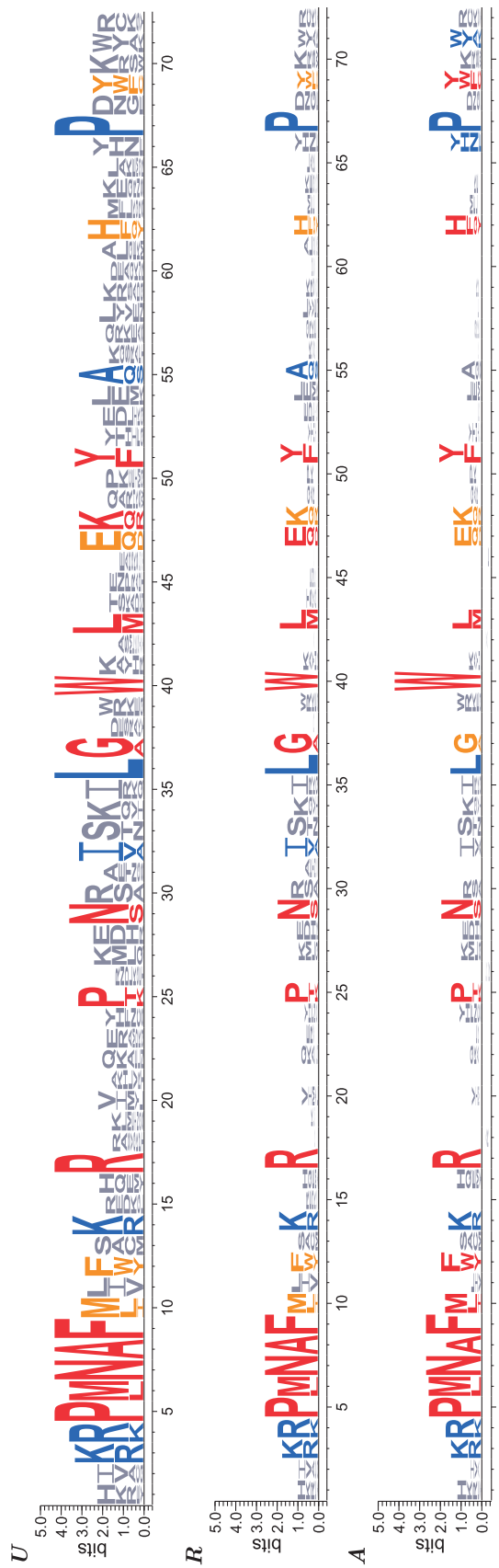

Fig. 1. Logos for HMGB box domains produced using the column information measures *U*, *R*, *A* and *B*. The alignment on which the logos are based is given in Supplementary Table S3. Based on a variety of considerations, a set of 18 positions are identified as 'important'. Specifically, from the structure of L-shaped NHP6A, positions 8, 12, 40, and 43 have been identified as primary hydrophobic core residues at the bend or crux, and positions 47, 48 and 51 as residues of 'alpha helix 3' to which they join, as illustrated in panel A of Supplementary Figure S1. Positions 17 and 25 are key residues maintaining the contacts between alpha helix 1 and 2 by hydrophobic interactions (Supplementary Fig. S1, panel B). These constitute all the interactions described in maintaining the compact structure of NHP6A (Masse *et al.*, 2002). For the DNA-binding contacts in NHP6A, positions equivalent to our 6, 9, 10 and 29 form hydrophobic wedges into the minor groove of the DNA (Masse *et al.*, 2002), (Supplementary Fig. S1, panel C). In addition, positions 5, 7, 37, 62 and 69 have previously been identified as important from an alignment of the larger family from which our subset of sequences are drawn (Landsman and Bustin, 1993), (Supplementary Fig. S1, panel D). For each measure *U*, *R*, *A* and *B*, columns colored blue identify the seven highest scoring 'false positive' positions, i.e. those not considered important by the criteria above. In contrast, columns colored red or orange identify important positions. Using the score cutoffs that yield seven false-positives for the various methods, red columns are 'true positives' and orange columns are 'false negatives'. Of the 18 important columns, the measures *U*, *R*, *A* and *B* yield 13, 13, 15 and 18 true-positives, respectively. The column colors shown here are for illustrative purposes only. The WebLogo program (Crooks *et al.*, 2004) colors columns based on amino acids properties, and our program inherits this coloring scheme. In Supplementary Figure S2, we show the HMGB box logos produced using this default coloring system

Table 1. Mean bit scores for catalytic and other sites, with standard deviations

Measure	Catalytic sites	Other sites
U	3.83 ± 0.71	2.27 ± 1.09
R	2.44 ± 0.74	1.08 ± 0.76
A	2.40 ± 0.96	0.76 ± 0.82
B	2.89 ± 1.09	1.02 ± 0.93

3.2 Catalytic sites

There are many structural and functional reasons for particular sites within protein families to be conserved, and it is difficult to systematically identify those that should be of most interest to a biologist. However, one class of positions that may be presumed to be of particular interest are catalytic sites within enzymes. Capra and Singh (2007) constructed a dataset of 660 multiple alignments of enzyme families in which the catalytic sites are known, and we have applied the measures U, R, A and B to this dataset, always first weighting the raw amino acid counts in each column. In Table 1, we show the resulting mean bit scores, and standard deviations for the dataset's 1999 catalytic sites, and 231 059 other sites.

We note, first, that when moving from measure U to R, the larger decrease in mean score among catalytic vis-a-vis other sites is caused by our weighting scheme (Altschul *et al.*, 2009; Sunyaev *et al.*, 1999) assigning on average fewer aggregate independent counts N to highly conserved columns. As with the HMGB box proteins, measure A yields even greater average corrections, but it also adjusts columns' scores with reference to the background frequencies $\vec{p}$. Measure B increases the mean bit score for all columns vis-a-vis A, but does so to a greater degree for the catalytic sites. In Figure 2, we compare distributions of the various measures' catalytic site scores after they have been normalized using the means and standard deviations of the non-catalytic site scores. For U, R, A and B, the mean-scaled catalytic site scores are, respectively, 1.43, 1.80, 2.01 and 2.02, and the median-scaled scores are 1.89, 1.96, 1.99 and 2.12. Thus, in comparison with U and R, the log-odds measures improve the separation of most catalytic site scores from those for the mass of non-catalytic sites.

Receiver-operating characteristic (ROC) analysis (Gribskov and Robinson, 1996; Schäffer *et al.*, 2001) is often useful for comparing retrieval methods. The relevance of this analysis is somewhat vitiated here, because the non-catalytic sites contain a large percentage of functionally important positions that have as much claim to be considered true-positives as the catalytic sites, and these sites are presumably heavily represented among those with highest score. Nevertheless, in Figure 3, we show $ROC_{35\,000}$ curves and scores for the pooled results from the catalytic site dataset, representing ~15% of all dataset columns. Measure U gives the uniform score of $\log_2 20 = 4.32$ bits to all completely conserved columns, and always prefers them to columns in which more than one type of amino acid is observed; this yields the long straight-line segment in the curve for measure U. Failing to correct for small sample size is a particular handicap when scores are given significance outside the context of an individual multiple alignment, because a highly conserved but non-uniform column from a large alignment may well be of greater interest than a uniform column from a small alignment: measure R thus easily outperforms measure U. By this ROC

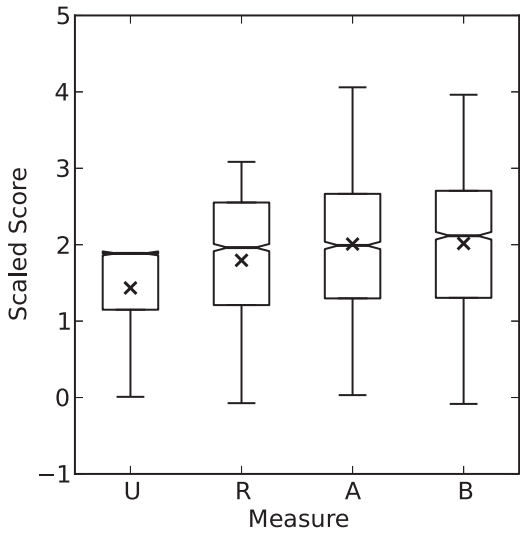

Fig. 2. Distributions of scaled catalytic site scores for the measures U, R, A and B. Raw scores are normalized by first subtracting the mean non-catalytic site score, and then dividing it by the non-catalytic site standard deviation. Boxes show the median and central quartile ranges for all scaled catalytic site scores. Error bars or 'whiskers' show the 5th to 95th percentile ranges. Mean-scaled scores are shown with the symbol 'X'

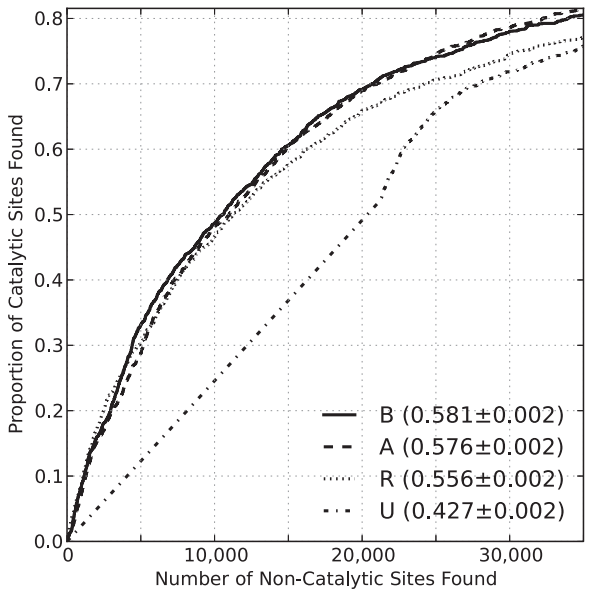

Fig. 3. $ROC_{35\,000}$ curves and scores for the measures U, R, A and B, calculated from the pooled column scores from the enzyme alignment dataset. The 1999 aggregate catalytic sites are considered true-positives or false-negatives, whereas the 231 059 aggregate non-catalytic sites are considered false-positives or true-negatives. $ROC_{35\,000}$ scores for all measures are given within the figure, along with standard deviations calculated as described in Schäffer *et al.* (2001)

analysis, measures A and B outperform R, but neither is significantly preferred to the other. Adding prior information concerning amino acid relationships increases, on average, all bit scores, as seen in Table 1, but it does not appear to favor catalytic sites more than highly conserved non-catalytic sites.

4 DISCUSSION AND CONCLUSION

We have proposed two distinct definitions for column information. Both are per-observation log-odds scores, taking the form $\frac{1}{N}\log_2[Q(\vec{x})/P(\vec{x})]$, where $Q(\vec{x})$ and $P(\vec{x})$ are probabilities for the column's data under alternative models of relatedness and chance. The first measure A calculates $Q(\vec{x})$ using the maximum-likelihood multinomial $\vec{q}=\vec{n}/N$ but then normalizes over the space of all possible observation vectors $\vec{x}$, so that Q is a probability. A is closest in spirit and form to the original measure proposed by Schneider and Stephens (1990). The second measure B derives $Q(\vec{x})$ from a Dirichlet mixture prior over multinomial space. B is the per-observation BILD score from Altschul *et al*. (2010).

In the DNA context, the assumption of ignorance underlying the measure A may yield the best information measure from the perspective of a molecule that is trying to recognize a DNA control element (Schneider, 1994). Other perspectives, however, are possible. Empirical study has shown that the multinomials characteristic of naturally occurring DNA motifs are more likely to be skewed toward a single nucleotide than even the Jeffreys prior would suggest. Thus, from a human perspective, more information for distinguishing motif from non-motif positions is attainable by using, e.g. a prior in which all $\alpha_j=0.25$ (Nishida *et al*., 2009).

When protein motifs are described with logos, even the molecular perspective of Schneider (1994) may be misleading. A protein may be seen as trying to achieve a particular conformation or structure within the constraints imposed by protein physics. The resulting correlations among amino acid frequencies at individual positions, which we discover empirically (Brown *et al*., 1993; Nguyen *et al*., 2013; Sjölander *et al*., 1996), may be taken as implicitly known to evolution. To measure the information inherent in a protein motif, it is thus reasonable to use a Dirichlet mixture prior that captures these correlations, and not to be bound by an uninformative prior.

All local pair-wise alignment scores may be seen as log-odds scores (Altschul, 1991; Karlin and Altschul, 1990), and all popular local alignment programs (Altschul *et al*., 1990, 1997; Pearson and Lipman, 1988; Smith and Waterman, 1981) may be understood as optimizing pair-wise log-odds scores. Extending the log-odds formalism to multiple alignments (Altschul *et al*., 2010) retains many of the features that render it useful for pair-wise alignment.

We have modified the open-source program of Crooks *et al*. (2004) to use the log-odds measures $A(\vec{x})$ and $B(\vec{x})$ as well as the entropy difference measures $U(\vec{x})$ and $R(\vec{x})$. Applying all these measures to an alignment of HMGB box proteins, and to a dataset of enzyme alignments, we found that log-odds scores facilitate the recognition of structurally and functionally important sites. For protein alignments, we prefer $B(\vec{x})$ on the theoretical grounds that prior knowledge concerning amino acids relationships is implicitly available to evolution and thus relevant to analyzing protein families. There is also some evidence that exploiting this knowledge increases the separation by score of biologically important from less important sites. We have made Web-based and stand-alone versions of our program available.

Funding: Y.K.Y., A.S., D.L. and S.F.A. were supported by the Intramural Research Program of the National Library of Medicine of the National Institutes of Health; J.A.C. was supported by institutional funds from Vanderbilt University.

Conflict of interest: none declared.

REFERENCES

Altschul,S.F. (1991) Amino acid substitution matrices from an information theoretic perspective. *J. Mol. Biol.*, **219**, 555–565.

Altschul,S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Altschul,S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Altschul,S.F. *et al.* (2009) PSI-BLAST pseudocounts and the minimum description length principle. *Nucleic Acids Res.*, **37**, 815–824.

Altschul,S.F. *et al.* (2010) The construction and use of log-odds substitution scores for multiple sequence alignment. *PLoS Comp. Biol.*, **6**, e1000852.

Altschul,S.F. *et al.* (1989) Weights for data related by a tree. *J. Mol. Biol.*, **207**, 647–653.

Brown,D.P. *et al.* (2007) Automated protein subfamily identification and classification. *PLoS Comput. Biol.*, **3**, e160.

Brown,M. *et al.* (1993) Using Dirichlet mixture priors to derive hidden Markov models for protein families. In: Hunter,L. *et al.* (eds) *Proceedings of First International Conference on Intelligent System for Molecular Biology*. AAAI Press, Menlo Park, CA, pp. 47–55.

Capra,J.A. and Singh,M. (2007) Predicting functionally important residues from sequence conservation. *Bioinformatics*, **23**, 1875–1882.

Colaert,N. *et al.* (2009) Improved visualization of protein consensus sequences by iceLogo. *Nat. Methods*, **6**, 786–787.

Cover,T.M. and Thomas,J.A. (1991) *Elements of Information Theory*. Wiley, New York, NY.

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

Gribskov,M. and Robinson,N.L. (1996) Use of receiver operating characteristic (ROC) analysis to evaluate sequence matching. *Comput. Chem.*, **20**, 25–33.

Grünwald,P.D. (2007) *The Minimum Description Length Principle*. MIT Press, Cambridge, MA.

Henikoff,S. and Henikoff,J.G. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA*, **89**, 10915–10919.

Henikoff,S. and Henikoff,J.G. (1994) Position-based sequence weights. *J. Mol. Biol.*, **243**, 574–578.

Jeffreys,H. (1946) An invariant form of the prior probability in estimation problems. *Proc. R. Soc. London Ser. A*, **186**, 453–461.

Karlin,S. and Altschul,S.F. (1990) Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl Acad. Sci. USA*, **87**, 2264–2268.

Landsman,D. and Bustin,M. (1993) A signature for the HMG-1 box DNA-binding proteins. *Bioessays*, **15**, 539–546.

Lawrence,C.E. *et al.* (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, **262**, 208–214.

Marchler-Bauer,A. *et al.* (2013) CDD: conserved domains and protein three-dimensional structure. *Nucleic Acids Res.*, **41**, D348–D352.

Masse,J.E. *et al.* (2002) The *S. cerevisiae* architectural HMGB protein NHP6A complexed with DNA: DNA and protein conformational changes upon binding. *J. Mol. Biol.*, **323**, 263–284.

Nguyen,V.A. *et al.* (2013) Dirichlet mixtures, the Dirichlet process, and the structure of protein space. *J. Comput. Biol.*, **20**, 1–18.

Nishida,K. *et al.* (2009) Pseudocounts for transcription factor binding sites. *Nucleic Acids Res.*, **37**, 939–944.

O'Shea,J.P. *et al.* (2013) pLogo: a probabilistic approach to visualizing sequence motifs. *Nat. Methods*, **10**, 1211–1212.
Pearson,W.R. and Lipman,D.J. (1988) Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.
Robinson,A.B. and Robinson,L.R. (1991) Distribution of glutamine and asparagine residues and their near neighbors in peptides and proteins. *Proc. Natl Acad. Sci. USA*, **88**, 8880–8884.
Schäffer,A.A. *et al.* (2001) Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res.*, **29**, 2994–3005.
Schneider,T.D. (1994) Sequence logos, machine/channel capacity, Maxwell's demon, and molecular computers: a review of the theory of molecular machines. *Nanotechnology*, **5**, 1–18.
Schneider,T.D. and Stephens,R.M. (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.*, **18**, 6097–6100.
Schneider,T.D. *et al.* (1986) Information content of binding sites on nucleotide sequences. *J. Mol. Biol.*, **188**, 415–431.
Schuster-Böckler,B. *et al.* (2004) HMM Logos for visualization of protein families. *BMC Bioinformatics*, **5**, 7.
Schwartz,R.M. and Dayhoff,M.O. (1978) Matrices for detecting distant relationships. In: Dayhoff,M.O. (ed.) *Atlas of Protein Sequence and Structure*. Vol. 5, (suppl. 3), National Biomedical Research Foundation, Washington, DC, pp. 353–358.
Sjölander,K. *et al.* (1996) Dirichlet mixtures: a method for improved detection of weak but significant protein sequence homology. *Comput. Appl. Biosci.*, **12**, 327–345.
Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.
Stormo,G.D. (1998) Information content and free energy in DNA–protein interactions. *J. Theor. Biol.*, **195**, 135–137.
Stros,M. (2010) HMGB proteins: interactions with DNA and chromatin. *Biochim. Biophys. Acta*, **1799**, 101–113.
Sunyaev,S.R. *et al.* (1999) PSIC: profile extraction from sequence alignments with position-specific counts of independent observations. *Protein Eng.*, **12**, 387–394.
Vacic,V. *et al.* (2006) Two Sample Logo: a graphical representation of the differences between two sets of sequence alignments. *Bioinformatics*, **22**, 1536–1537.
Workman,C.T. *et al.* (2005) enoLOGOS: a versatile web tool for energy normalized sequence logos. *Nucleic Acids Res.*, **33**, W389–W392.