Compensation for nucleotide bias in a genome by representation as a discrete channel with noise

Mark Schreiber and Chris Brown*

Department of Biochemistry, University of Otago, PO Box 56, Dunedin, New Zealand

Received on August 28, 2001; revised on October 31, 2001; accepted on November 2, 2001

Abstract
Motivation: Calculation of the information content of motifs in genomes highly biased in nucleotide composition is likely to lead to overestimates of the amount of useful information in the motif. Calculating relative information can compensate for biases, however the resulting information content is the amount seen by an observer and not by a macromolecule binding to the motif. The latter is needed to calculate the discriminatory power of the motif and to compare motifs between species.

Results: By treating a biased genome as a discrete channel with noise, in accordance with Shannon Information Theory, we were able to remove both 'Distortion' and 'Noise' from the motif and recover a more instructive biological 'signal.' A Java application, LogoPaint, was developed to remove nucleotide bias distortion and triplet frequency noise from motifs, calculate information content and present the motif as a logo. We demonstrate how this technique can 'unmask' motifs in the translation initiation regions of bacteria that are obscured by strong sequence biases.

Availability: LogoPaint is available to all users from the authors as an executable JAR file. Source code is available by arrangement.

Contact: chris.brown@stonebow.otago.ac.nz

Introduction

In 1948, Shannon (Shannon, 1948) introduced a mathematical theory of communication that described the communication of symbols through a channel. Later this theory was extended to the study of nucleic acid motifs and their macromolecular recognizers (Schneider et al., 1986; Schneider and Stormo, 1989; Schneider, 1991a,b). This allowed the calculation of the amount of information contributed by individual nucleotides in the motif. One could also calculate if the information gained by a molecular machine upon binding to a motif would be sufficient to discriminate that motif from among all possible binding sites (Schneider et al., 1986).

Shannon proposed that an encoding system could be designed that would allow errors to be reduced to a desired level as long as the capacity of the communications channel is not exceeded. The information in DNA and RNA can be encoded using four symbols, however application of Shannon’s theory to such an encoding scheme can lead to overestimates of the amount of useful information if the symbols are not observed with approximately equal frequency.

In a genome where the four nucleotides are not equally distributed, motifs appear to show conservation of the over represented residues. However, the conservation is also found elsewhere in the genome suggesting this information is not useful in recognizing the motif. One solution is to calculate the information content of the motif relative to a background distribution (Stormo, 1998). This technique is useful for determining the difference between the observed motif and the background distribution but as Schneider (1999) points out relative information is not a state function. Because of this it cannot be used to infer the amount of information gained, or the amount of energy dissipated, upon recognition of a motif by a macromolecule, making comparisons of relative information content within and between organisms impossible.

We propose that a genome with a skewed nucleotide distribution is best viewed as a discrete channel with noise. Shannon’s 1948 paper described a case where ‘a signal is perturbed by noise during transmission at one or other of the terminals.’ In this way, a received signal may differ from the transmitted signal. The statistically ideal genome would contain all four nucleotide residues at equal frequencies. Through mutational processes and selection pressure, the genome of an organism may not meet this statistical ideal. An unbiased analysis of motifs in a skewed genome requires that the transmitted signal must be recovered. To recover the original signal the ‘noise’ must be removed from the genome.

Shannon described two types of perturbations, distortion and noise. If a transmitted signal S is received E with a constant bias then the function f that describes the bias is known as the distortion function. Distortion is described

*To whom correspondence should be addressed.
by the equation:

\[ E = f(S). \]  

(1)

This process is equivalent to mutational biases in the replication of DNA leading to a general genome-wide bias. Distortion may be corrected if the function \( f \) can be determined (or closely estimated) and the inverse of that function applied to the received signal. Rearranging (1), we can derive an equation that reproduces the original signal:

\[ S = f(E)^{-1}. \]  

(2)

Noise is a stochastic process where a received signal does not always undergo a uniform distortion function. Noise can be modelled as a Markov chain where for a finite number of states there is a probability that if the transmitter is in state \( \alpha \) while emitting symbol \( i \), that symbol \( j \) will be received leaving the transmitter in state \( \beta \).

\[ p_{\alpha,i}(\beta,j). \]  

(3)

A state is a genome or genome region with common statistical characteristics. In a genome there may be a number of states that are particularly ‘noisy’ including mutational hotspots, sequences that cause polymerase slippage, and coding regions where triplet biases are present. Noise in these regions is not uniform and is state dependant. Algorithms of greater complexity are needed to reduce or remove this kind of noise.

During the evolution of an organism, mutational processes can introduce biases in the genome. The macromolecules that recognize motifs in these genomes co-evolve to recognize motifs in the presence of these biases. By evolving the ability to ignore these biases, the recognizer molecule has effectively evolved a noise or distortion correction function.

**MATERIALS AND METHODS**

**Materials**

Bacterial translation initiation and termination contexts were extracted from the TransTerm database (Jacobs et al., 2000). Initiation contexts were aligned to the start codon and trimmed to 25 bases upstream and 21 bases downstream of the 3 base start codon to give a ribosome footprint of 49 nucleotides. Nucleotide biases were corrected using the nucleotide composition observed 100 bases upstream and downstream of the annotated start codon except in the case of triplet noise, which was corrected using the observed frequency of each nucleotide at each position of the codon. Because codon usage at the 5’ end of a gene often differs from that found later in a gene only the first 99 bases of all annotated coding regions were used to calculate triplet frequencies (data extracted from TransTerm). Predicted gene numbers, gene locations and genome sizes where extracted from TIGR’s Comprehensive Microbial Resource (Peterson et al., 2001; www.tigr.org/tigrscripts/CMR2/CMRHomePage.spl).

**Algorithms**

For a uniform bias or distortion there is only one state, its distribution can be calculated by sampling the genome region modelled by that state. Distortion may be corrected by calculating the likely frequencies of each nucleotide at each position of the motif in the ‘statistically ideal’ genome using the formula:

\[ n’_i = P_i \frac{P(i | S)}{P(i | \alpha)} \sum n_i; \]  

(4)

where \( n’_i \) is the corrected number of \( i \) nucleotides, \( P_i \) is the observed frequency of \( i \) \( P(i | S) \) is the probability of seeing nucleotide \( i \) given the statistically ideal genome \( S \) \( (0.25) \) and \( P(i | \alpha) \) is the probability of nucleotide \( i \) being emitted by state \( \alpha \).

Equation (4) can be rearranged to give:

\[ \frac{n’_i}{\sum n_i} = \frac{P_i}{P_i^{(\text{observed})}} \quad \]  

(5)

where \( \frac{n’_i}{\sum n_i} \) is the desired new frequency \( f’_i \), \( P(i | S) \) is the ideal probability \( P_i^{(\text{ideal})} \) and \( P(i | \alpha) \) is the observed probability \( P_i^{(\text{observed})} \). Note that when dealing with genomic data the observed and ideal probabilities can be thought of as frequencies. Thus, (5) becomes:

\[ f’_i = f_i \frac{f_i^{(\text{ideal})}}{f_i^{(\text{observed})}}. \]  

(6)

Many amino acids are redundantly coded by more than one codon. The set of codons that code for one amino acid are generally conserved at the first two positions and variable at the third ‘wobble’ base (Crick, 1966). In some organisms, where there is a choice of codons for an amino acid, there is a selection for codons ending in particular residues, often G or C (Karlin et al., 1998). This produces a three-nucleotide periodic distortion in the coding region. In the case of triplet noise there are three observed states, \( a_k \) where \( k = 0 \cdots 2 \) (numbering from zero), in the coding region (and (4) may be generalized to:

\[ n’_i | k = P_i \frac{P(i | S)}{P(i | a_k)} \sum n_i. \]  

(7)

The above equations describe a linear correction for distortion. We believe a linear function is best as it most closely follows the stepwise nature of evolutionary change, which generates the observed noise. The LogoPaint program offers the opportunity to make linear distortion corrections as well as corrections based on a square of the ideal to
the observed (8) or a correction that increases its influence
with increasing distortion (9).

$$f'_i = f_i \left( \frac{f_i(\text{ideal})}{f_i(\text{observed})} \right)^2, \quad (8)$$

$$f'_i = f_i \left( \frac{f_i(\text{ideal})}{f_i(\text{observed})} \right)^{1+\ln \left( \frac{f_i(\text{ideal})}{f_i(\text{observed})} \right)}, \quad (9)$$

A correction based on (8) would imply that the distortion of a motif increases rapidly once distortion has begun, effectively a speeding up of the evolutionary process. A correction based on (9) would imply that the distortion of the recognizers target proceeds nearly linearly at first but begins to increase rapidly once a certain threshold is reached. For biases typically observed in genomes, (9) roughly approximates the linear correction. Although we believe the linear correction is generally applicable, there may be exceptional circumstances where these other corrections are more appropriate. Models of greater complexity could be designed to correct for other biases such as triplet expansions or mutations introduced by polymerase slippage.

After the application of the correction algorithm, we calculated the information content of the motifs using Shannon’s uncertainty measure:

$$H = -\sum_i P_i \log_2 P_i, \quad i \in \{a, c, g, t\}. \quad (10)$$

Because a state function is used to calculate the information of a motif, the free energy of interaction between the motif and its recognizing macromolecule can be calculated.

The amount of information $R_{freq}$ required to discriminate a binding site occurring $\gamma$ times in a genome of size $G$ was calculated using the formula of Schneider et al. (1986),

$$R_{freq} = -\log_2 \frac{\gamma}{G} \text{ (bits per site).} \quad (11)$$

For the following analysis we have used samples of over 855 sequences thus corrections for small samples can be ignored however an unexplored aspect of the present algorithm is how it relates to the small samples.

**Implementation**

The genome distortion and triplet noise reversal algorithms were implemented in Java 1.3 and incorporated into a DNA logo-generating program, LogoPaint. LogoPaint makes use of the Java Advanced Image library (JAI 1.1) (http://java.sun.com/products/java-media/jai/) to generate logos in BMP, JPEG and TIFF. LogoPaint also uses and extends some classes from the Biojava open source project (http://www.biojava.org) to draw the logo and calculate the information content of a motif. LogoPaint is available to all users from the authors as an executable JAR file. Source code is available by arrangement.

**RESULTS AND DISCUSSION**

**Results**

To demonstrate the utility of the algorithm we analyzed motifs from the translation initiation contexts of bacterial genomes. The genome of *Borrelia burgdorferi* (Fraser et al., 1997) has an extreme A/U content bias in the region flanking the start codons, (39.5% A, 10.3% C, 15.0% G, 35.2% U). The translation initiation context of *B. burgdorferi* (Figure 1a) contains 17.0 bits of information but a 910.7 kb genome with 855 translation initiation sites requires only 10.1 bits of conserved information at these sites (Schneider et al., 1986). Possibly the initiation context may contain other motifs that require the remaining 6.9 bits but more likely the A/U genome bias is causing an artificially high level of conservation. By applying our correction to the genome distortion a more realistic figure of 7.0 bits was obtained indicating that the majority, but possibly not all, the A/U bias is ignored by the ribosome. The characteristic purine rich Shine–Dalgarno (SD) sequence found at most translation initiation sites is obscured in the uncorrected motif but is clearly visible in the corrected (Figure 1b). Slight flexibility in the position of the SD, which is not captured by this alignment to the start codon, is a possible reason for the lower than expected information content of the motif.

After correction of the *Borrelia* logo more subtle features became apparent including a slight purine bias at the −4 position and codon biases downstream of the start codon, *Escherichia coli* also contains codon biases and a slight purine bias at the −3 position (Figure 1c). The corrected *Borrelia* logo (Figure 1b) also indicates that a GUG start codon is preferable to a UUG start codon whereas the uncorrected logo suggests the opposite. There is conflicting evidence as to which is biologically preferred in *E. coli*. The evidence of Weyens et al. (1985) and Sacerdot et al. (1996) suggest that GUG is the better start codon. GUG is also used by more genes than UUG. The experiments of Stenström et al. (2001a) and Sussman et al. (1996) indicate that UUG is a slightly better initiation codon. The differences between the two initiation codons are small and the presence or absence of other motifs may cause one to be stronger than the other in certain circumstances. Similar situations are likely to occur in other organisms.

As a control, we applied our distortion correction to *E. coli*, a species with a largely unbiased nucleotide distribution in the region surrounding its start codons (26.8% A, 22.6% C, 24.1% G, 26.6% U). Here the correction had very little effect (Figure 1d versus c).
Fig. 1. (a) An RNA logo of the initiation context of *B. burgdorferi* showing the conserved start codon. All other motifs are obscured by the extreme A/U distortion. The high frequency of U nucleotides at the first position of the start codon (position 1) is most likely caused by nucleotide distribution bias. (b) An RNA logo of the initiation context of *B. burgdorferi* corrected for A/U distortion. The G/A rich SD is now clearly visible. The first position of the initiation codon now shows that A and G convey more information to the initiation machinery which would be expected from comparative biology to other microorganisms. (c) An RNA logo of the uncorrected initiation context of *E. coli*. (d) An RNA logo of the initiation context of *E. coli* corrected for nucleotide frequency distortion. (e) An RNA logo of the initiation context of *P. aeruginosa*. The purine rich SD is clearly visible as a large G/A peak approx. 9 bases before the start codon. (f) An RNA logo of the initiation context of *P. aeruginosa* corrected for nucleotide distortion. Despite the G/C bias of the genome the contribution of Gs to the SD is still clear in the corrected motif.
Interestingly the initiation context of *E. coli* (Figure 1d) also contains slightly less information (7.2 bits) than required (10 bits) which may also be due to flexibility in the position of the SD.

We also tested the ability of the correction technique to maintain a G rich motif in the presence of a G rich context. Figures 1e and f show the logos of the translation initiation contexts of *Pseudomonas aeruginosa* uncorrected and corrected respectively. Despite the high G/C bias (18.3% A, 33.6% C, 29.9%G, 18.2%U), the contribution of Gs to the SD was still clear in the corrected motif although the influence of the conserved G residues was slightly reduced.

To demonstrate the correction for triplet noise we examined the region immediately downstream of the initiation codon of *E. coli*. The presence of a coding region will introduce nucleotide biases due to favouring of one or more nucleotides at certain positions in the codon. Figures 2a and b show the uncorrected and corrected logos respectively. With the triplet noise removed, a subtle bias towards the use of As was seen. Stenström *et al.* (2001a,b) demonstrated that codons downstream of the start codon containing two or more adenosine residues favour enhanced translational efficiency in *E. coli*.

**Discussion**

The unmasking of subtle motifs and the more realistic information contents obtained in these studies suggest that applying noise and distortion reversal functions to motifs in biased contexts is a useful and valid technique. The results also lend some weight to the possibility of evolved noise tolerance in molecular machines. Most likely the ‘correction’ would take the form of modified sensitivity to the presence of certain residues in a motif.

Comparisons of macromolecules that bind similar motifs in different noise backgrounds may reveal the physical attributes that allow them to tolerate noise and distortion. An example may be drawn from the 16S rRNA from *Borrelia*. The SD binding sequence at the 3' end of the 16S rRNA has the same sequence as in *E. coli* suggesting that an *E. coli*-like SD would be bound. However, due to the extreme genome bias in *Borrelia* an *E. coli*-like SD motif is obscured, suggesting that other structures in the 16S rRNA or in the proteins of the ribosome increase its sensitivity to the ‘expected’ residues and reduce its sensitivity to the background noise.

The information encoded by a distorted background distribution can be considerable, almost 10 bits in the case of the *Borrelia* ribosome binding sites. If the information were transferred to the binding molecule it would need to be later lost as entropy. Given that a considerable amount of this information does not help the molecule discriminate the correct site, molecules that evolve noise tolerance strategies could be significantly more efficient.

The algorithms presented here allow the assessment of motifs in any distorted or noisy genomic background where the source of the distortion or noise can be modelled or estimated. Because Shannon’s state function is used to calculate the corrected amount of information calculations can determine the amount of information gained by a macromolecule upon binding a motif and the amount of entropy that must be dissipated by the binding molecule. Comparisons of corrected information content may also be made between species.

**ACKNOWLEDGEMENTS**

This work was partially supported by a University of Otago targeted research scholarship. Thank you to the anonymous reviewers for their comments.

**REFERENCES**


