Additional file 2

Additional text – Relationship between $R_{iw}(b,l)$ and $R_{sequence}(l)$:

The individual information content of each individual splice site was calculated using the following equation [29, 30]:

$$R_{sequence}(l) = 2 + \sum_{b \in A,C,G,T} f(b,l) \log_2 f(b,l) \qquad (2)$$

where, $f(b,l)$ is the probability of nucleotide b at position l. To calculate the information content $(R_i, bits)$ of each splice site, at first, an individual information weight matrix is generated from the frequencies of each nucleotide at each position. The individual information weight matrix $R_{i,w}(b,l)$ can be calculated from the following equation [29, 30]:

$$R_{i,w}(b,l) = 2 + \log_2 f(b,l) \qquad (3)$$

The information content of each splice site was calculated by summing up $R_{i,w}(b,l)$ at each position of the splice site sequences. The relationship between $R_{i,w}(b,l)$ and $R_{sequence}(l)$ is discussed below.

In a set of aligned sequences the jth sequence is represented by a matrix $s(b,l,j)$, where b is the base at position l. The individual information of sequence j is the dot product between the sequence and the weight matrix $R_{i,w}(b,l)$ as given below:

$$R_i(j) = \sum_l \sum_{b=A}^{T} s(b,l,j) R_{iw}(b,l) \qquad (4)$$

The frequency matrix $f(b,l)$ is created by aligning n individual sequences and given by:

$$f(b,l) = \frac{1}{n} \sum_{j=1}^{n} s(b,l,j) \qquad (5)$$

As the frequencies also sum to one, we can write:

$$\sum_{b=A}^{T} f(b,l) = 1 \qquad (6)$$

The mean information content of n sequences used to create the frequency matrix $f(b,l)$ is given by:

$$E(R_i) = \frac{1}{n} \sum_{j=1}^{n} R_i(j) \qquad (7)$$

From equation (3) and (4) we get:

$$R_i(j) = \sum_l \sum_{b=A}^{T} s(b,l,j)[2 + \log_2 f(b,l)] \qquad (8)$$

From equation (7) and (8) we get:

$$E(R_i) = \frac{1}{n}\sum_{j=1}^{n}\sum_{l}\sum_{b=A}^{T} s(b,l,j)[2+\log_2 f(b,l)] \qquad (9)$$

From equation (5) and (9) we get:

$$E(R_i) = \sum_{l}\sum_{b=A}^{T} f(b,l)[2+\log_2 f(b,l)] \qquad (10)$$

$$= \sum_{l}\sum_{b=A}^{T} 2f(b,l) + \sum_{l}\sum_{b=A}^{T} f(b,l)\log_2 f(b,l)]$$

From equation (6) and (10) we get:

$$E(R_i) = \sum_{l} 2 + \sum_{l}\sum_{b=A}^{T} f(b,l)\log_2 f(b,l)] \qquad (11)$$

The right hand side of equation (11) is similar to that of $R_{sequence}(l)$ in equation (2). Hence, it can be deduced that the average of individual information content is the average information content of the sites.