# 1 Extraction Details

For each position, $i$, we compute the KL-Divergence between the observation distribution of the match state and the background distribution.

$$h_i = \sum_{x \in \text{Symbols}} m_{ix} * \log(m_{ix}/b_x) \tag{1}$$

We then normalize this vector so that the values fall between 0 and 1: $h_i = h_i / \max_i(h_i)$.
Smoothing is done according to the following equation:

$$h_i^s = \frac{1}{s} \sum_{j=0}^{s-1} h_{i+j} \quad \text{for } i \in [0, H_l - s]. \tag{2}$$

So, for example, $h_0^s$ would be the average value of $h_0$ through $h_{s-1}$. Then the window shifts one position forward and $h_1^s$ would be the average value of that window. This continues up to $h_{H_l - s}^s$, where $H_l$ is the length of $h$.

Every subset of $h$ which lies above the average value of $h$ is excised. We then perform a clean up step which removes edge positions with low KL-Divergence. For each fragment, we first find the max value and then, starting from the beginning, remove every position with a value less than 50% of the max value, stopping at the first position for which this not true. We then do the same thing starting from the end of the fragment. This noise at the beginning and end is caused by the averaging method. Note that this clean up procedure can cause the fragment to become shorter than the given minimum length, but it will never produce a fragment with length 0.