Supplementary Figure 1: PhylCRM scoring scheme for a single motif

(a) $g$ represents the sequence being searched for CRMs and $a^1$ and $a^2$ are sequences from another organism aligned to it. $L$ represents the length of the sequence, $H_{n,j} = g$, $H_{n,j} = a^1(0)$, and $H_{n,j}$ denotes the alignment column at position $j$.

(b) Tree indicating the phylogeny of $g$, $a^1$, and $a^2$. (c) Scoring motif matches using the MEHB model. Here, the probability that a given nucleotide $a$ turns into $b$ during time $t$ is given by a matrix exponential, for a suitably chosen rate-matrix $R$. This probability is then used to compute the probability of observing the set of nucleotides $H_{n,j}$ under both the MEHB rate-matrix and the neutral matrix. The score of the motif $\phi$ is then taken to be the log-likelihood of the ratio of these probabilities. (d) Graphical image of scores for a motif $M$ along $g$, where the height of the bars is $\phi/m$. These scores are stored in an array $\xi$ and the score of a window $w_j$ (represented by $\Xi(w_j)$) is then given by summing $\xi$ in $w_j$. (e) When there is no alignable sequence at a given position (or if there is no motif match there), the branch containing that sequence is removed and the pruned tree is used to compute $\phi$. 