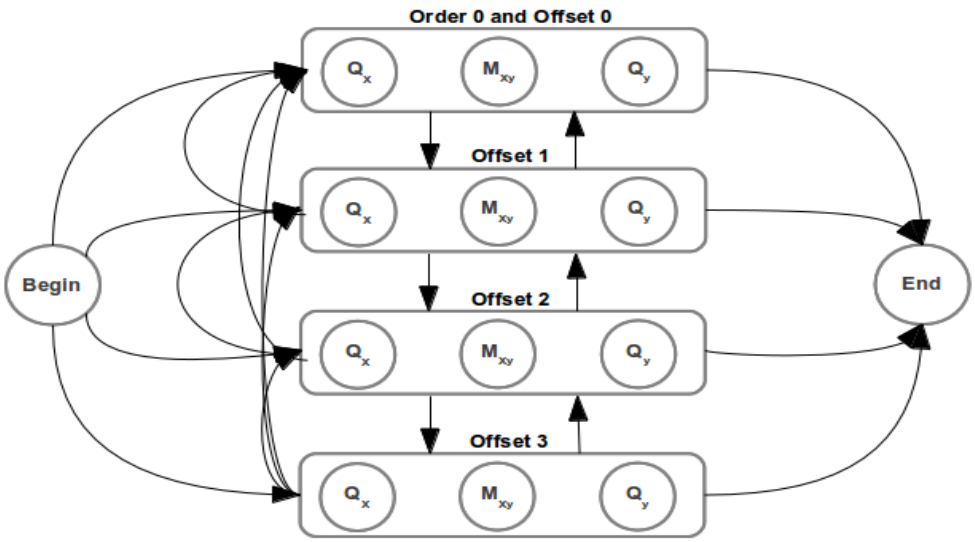
Utilizing probabilistic dependencies in low-complexity regions when detecting homologous regions of biological sequences

Thomas M. Poulsen, Martin C. Frith, and Paul Horton

Computational Biology Research Center (CBRC), AIST Tokyo Waterfront Bio-IT Research Building

The ability to analyze and find homologous regions between nucleotide sequences is a fundamental topic in biological research. One difficulty in detecting homologs is that they may contain tandem repeats and low complexity regions that produce alignments that are not homologous. Standard methods attempt to address this issue by masking such regions, but masking does not distinctively model the difference between alignment of low complexity and 'standard' regions, and thus risks disregarding important information. In this study, we employ probabilistic models to more accurately quantify the alignment of repeats and low-complexity regions. In particular, this involves the design and implementation of Hidden Markov based architectures (HMMs) for aligning sequences of homologous regions. A comparison is made between a standard pair HMM model and extended models that are able to detect repetitive regions while also quantifying higher order nucleotide distributions (see example of an extended model in figure below).




```
TGTGTGGATGGGTGTGTGTGGTGTGGTGAGTGGTATGTGGTGTGGTGTGTTGTGTGTTGTGTGGTGT---GT
TGTGTGGATGGGTGTGTGTGGTGTGTTGTGTGTTGTGTGGTG---TGTGTGGTGTGTTGTGTGGTGTATGGT
--02--02--02--02--02--02----03----03----03xxxmmm----03----03----03myyymm
```

Top: Example of an 'extended' pair HMM model that includes standard pair HMM states (shown under order-0 and offset-0) and offset states (in this example offsets 1, 2, and 3). Offset states function as first order Markov states but employ emission probabilities that are dependent on nucleotides *x* number of indexes prior to the current alignment position. For example, offset-1 is dependent on the nucleotide one index before the current position, offset-2 two indexes before, and offset-3 three indexes before (models can also incorporate higher offsets as well as higher orders). Offset or non-offset states can either produce alignments (*m* states) or insertions (*q* states) - not all connections shown. **Bottom:** Example of an alignment produced by the model shown above illustrates how the model enters the offset states throughout a number of repeats in the two sequences. The first two rows contain the actual alignment produced, while the last row displays the states that the model entered throughout a Viterbi alignment. In this example, '--O2' and '----O3' correspond to matches produced by states offset-2 and offset-3, while 'm', 'x', 'y' correspond to matches/insertions produced by order-0 and offset-0 states.