Gilles Didier (IML)

Variable length decoding of sequences

We present a generalization of the the so-called N-order local decoding of sequences, introduced in a previous work. This decoding consists in relabelling each position of a set of sequences, in a way depending on its environments of length N (*i.e.* the N sub-sequences of length N containing this position). In practice, the approach remains to label in the same way two positions i and j, if there is a subsequence of length N occurring both at i-k and j-k with k<N, or if there is a chain of positions from i to j, granting successively these property.

With the variable length decoding of sequences, the environments of each position are not of a same length N but vary according to the local sequence of nucleotides. The sequence of variable length environments of a sequence s is defined as the sequence of elements of a prefix code successively observed along s. The prefix property, that is there is no word of the code which is prefix of another, ensures that this sequence of environment is well defined. Two positions i and j are now decoded in the same way if there is a element of the prefix code of length greater than k, occurring both at i-k and j-k or again a transitive chain of this last relation between i and j.

We first give a algorithm computing this variable length local decoding of a set of sequences which runs in linear time and uses a linear memory space. The decoding is then applied to the comparison of sequences by deriving from it, the following distance between two sequences s and t: the number of common decoded symbols between s and t, normalized by the shortest length. The main difficulty is to define a prefix code somehow relevant with regard to a given set of sequences to compare. We construct this prefix code by considering the shortest subsequences verifying two properties: (1) they occur at most once in each sequence of the set; (2) their probability, under a Markov model, of occurring more than twice in the whole set is below a given threshold .

Next, we evaluate this approach over several viral genomes alignments (HIV, Dengue, Hepatitis, Influenza). For each alignment, we measure the correlation between the homology percentage, computed from the alignment, and our distance, obtained from the raw sequences included in the alignment. The distance obtained with the variable length decoding shows better results than a N-order local decoding (for any N), as well than another alignment-free distance based on Maximum Unique Match.

Finally, we investigate the ability of the variable length local decoding to define anchors for multiple alignment of sequences. To do this, we take the viral genomes alignments as references and measure how often two positions decoded in the same way belong to a same column of the alignment.