Abstract Guénoche :

In this paper, we compare the accuracy of four string distances on complete genomes to recover correct phylogenies using simulated and real biological data. These distances are based on common words shared by raw genomic sequences and do not require preliminary processing steps such as gene identification or sequence alignment.
Moreover, they are computable in linear time.

The first distance is called MSM, because it is based on Maximum Significant Matches. The second is the KW distance which is computed from the frequencies of all the words of length $k$ [Qi, {\it et~al}., 2004]. The third distance, called ACS, is based on the Average length of maximum Common Substrings in any position [Ulitsky {\it et~al}., 2006].
And the last distance ZL is based on the Ziv-Lemple [1977] compression algorithm.

We describe a simulation process of evolution to generate a set of sequences having evolved according to a random tree topology $T$.
This process allows both base substitutions and fragment insertion/deletion, including horizontal gene transfers.
The distances between the generated sequences are computed using the four string formulas and the corresponding trees $T'$ are reconstructed using Neighbor-Joining. $T$ and $T'$ are compared using three topological criteria, the Robinson-Fould distance, the percentage of quadruplets with different tree topology and the Maximum Agreement Subtree size. These comparisons show that the MSM distance outperforms the KW, ACS and ZL distances whatever the parameters used to generate sequences.

Finally we test the MSM distance on real biological data (i.e. prokaryotic complete genomes), we compare the trees based on NJ applied to the MSM distance to the ML reference tree of 16S + 23S RNA aligned sequences. We show that the MSM distance provides accurate results to study intra-phylum relationships.

This is joint work with C. Brochier and F. Guyon.