Substring Differences in Genomes

Hendrik Jan Hoogeboom Walter A. Kosters
Jeroen F. J. Laros
Leiden Institute of Advanced Computer Science
Universiteit Leiden, The Netherlands
jlaros@liacs.nl

December 22, 2008

Unique substrings in a genome are used for various purposes, with applications varying from the development of primers, applied in *PCR* or *MPLA* experiments, to the fabrication of *microarrays*. These unique strings can also be used for the *identification of DNA samples*. If a string is known to be present in one species, and not in the other, we can determine the origin of a sample by employing a carefully chosen mix of identifying substrings (markers). An extension of this idea is to utilise unique or rare strings as a basis for a *distance measure* (metric) for genomes. Such a metric can for example be the basis for the construction of a phylogenetic tree.

We introduce a new distance measure for genomes. This metric is based upon the occurrence of relatively small substrings (length 14 to 18) in both genomes. For each substring, we count the number of occurrences in each genome, yielding a matrix where at position (x, y) the number of strings found exactly x times in the first and y times in the second genome is recorded. For practical purposes, x and y both have the same maximum. If a certain string is found more than this maximum (typically 4 or 16), it occurs 'often'. Construction of this matrix can be done efficiently.

After this step, we have a large amount of markers, substrings that are present in one genome, but not in the other, that can be used for distinction between the two genomes. The size of this set adds to the reliability of the results.

We use the constructed matrix to derive a distance between the two genomes. For this purpose, we propose different metrics, defined on sets or *multisets*. First we adapt the well-known Jaccard distance. This metric is defined on sets alone, so strings that are present more than once are counted as one.

Secondly, a metric is designed specifically. It relies on a second matrix filled with weights to differentiate between $(0, 1)$ and $(0, 2)$ for example. The assumption that this difference is important arises from the fact that if a substring is not present in one genome and two times in the other, the string is probably more important than when it is only present once in the second genome.

The third metric used is one defined on multisets. The main advantage of this metric is that no weighing matrix is needed, since it already takes the difference

in occurrences into account. The way these differences are accentuated is done by the choice of a parameter function f, indicating the difference in occurrences of a substring used. The choice of f is quite intuitive and based upon domain knowledge. This function must represent the fact that shared occurrence is more important than the difference in the number of occurrences.
Various experiments with these metrics are performed and visualisations of the results are shown. We also provide concrete examples of functions that will perform well within the task described above.