Poster C-27

LOSS: an heuristic approach to incorporate prior information into prokaryotic genomic GC prediction

Authors:
Hugo Naya (*Unidad de Bioinformática, Institut Pasteur de Montevideo, Uruguay*)
Héctor Romero (*Laboratorio de Organización y Evolución del Genoma, Sección Biomatemética, Facultad de Ciencias, Uru*)

Short Abstract: Genomic GC content is one of the most important features of a genome. We introduce an heuristic algorithm that allow to incorporate information from completely sequenced genomes to improve the genomic GC content prediction from a sample of genes. In general, performance of LOSS algorithm compared favorably with previous methods.

Long Abstract:

Genomic GC content is one of the most important features of a genome. This is especially true among prokaryotes where, generally, the vast majority of the genome is composed of coding DNA. These organisms are characterized by an extremely broad compositional range, since GC levels are approximately comprised between 25% and 75%. Formerly determined from biophysical–biochemical methods (the most important being density gradient centrifugation and thermal denaturation midpoint methods), the genomic age shifted the possibilities for obtaining its real values directly from the sequences, which allowed the possibility of checking the consistency of previous methods. Previous works mentioned an ‘‘average accuracy’’ of ±2.7% in genomic GC content.

Although the number of completely sequenced genomes is increasing, there are numerous genes from genomes that do not appear as targets for sequencing projects in the near future. Furthermore, recent studies focused on genomic differences between ecological niches based in massive sequencing, which render a high number of sequences available but few completed genomes in the near horizon.

Recently, Zavala et al. (2005) showed that the genic sample mean is a consistent estimator of the genomic GC. With a random sample of 10 or more genes, the estimates obtained by this method are generally better than reported from biophysical–biochemical methods.

However, available sequences are clearly never a random sample of the genic pool of each organism. In such cases, estimates from the sample mean could become highly biased. With the aim to overcome these caveats and to reduce, if possible, the sampling variance of the “mean” method we introduce here a general approach that allow to include “prior” information in the estimation process.

Organisms are evolutionary related and this relationship includes different genomic and genic features. When analyzing genes and GC content, is reasonable to suppose that, as more related two organisms are, genes with GC content under (or over) the mean in one genome tend to be also under (or over) the respective mean in the other. When the phylogenetic relationship decrease, the action of different evolutionary forces (e.g. selection and mutation) becomes important, tending to erase the previous signal.

The heuristic approach described here is based on this logic of sequence similarity (LOSS) and clearly it is only one possible implementation among several different. Given a sample of x genes from a prokaryotic organism with unknown genomic GC content and a set N of

completely sequenced genomes, the basic algorithm can be described as:

for each gene in x
perform a blastp search against N
pick the best hit for each genome in N
for each hit
calculate the difference D in GC between the corresponding genomic value and the hit
calculate the weighted average of differences (D) by a function of the blast results (e-values)
calculate a dispersion measure v within the gene
add the weighted difference (D) in GC to the gene to obtain gene_new
calculate the weighted average of gene_new by an inverse function of v

The rationale behind this algorithm consist in that orthologous genes can be identified by the blast search and that the conservation in the difference between the GC content respect to the genomic GC is inversely related to the phylogenetic distance (blast results). Further cautionary penalization is added to genes with very inconsistent or unexpected behaviour, that is, very big or extremely variable differences.

To evaluate the performance of this approach we selected fifteen organisms that represent major lineages in prokaryotic evolution. From the set of available completely sequenced organisms a selection was made in each case to match adequately a taxonomical level of the sample.

Evaluation was conducted in two different set of samples. For each of the fifteen selected genomes, one hundred random samples of size ten were taken. This represents the hypothesis tested in Zavala et al (2005) in which the “mean” method performed very well. A second set of fifty biased samples were taken for each genome between the one hundred genes with most extreme GC content at both sides. Performance was evaluated by “Mean Squared Error” (MSE), bias, variance and times LOSS over performed the “mean” prediction.

Results were highly dependent on the existence of related organisms within the set of sequenced genomes. This is especially true for Archaea where completely sequenced organisms are scarce. In these cases the performance of LOSS was forced to nearly identical to the “mean” method. In clades with more representatives LOSS usually performed better; by reduction of the sampling variance when samples are taken at random or by decrease in the bias in the other case.

In some organisms, with biased samples the behaviour of LOSS was asymmetric: estimates from samples with relatively high GC were improved while predictions from samples with low GC do not improve.

One interesting feature we observed is related to the fact that biased samples that usually do not improve by this method are associated with the absence of significant matches in blastp. This can help to mark suspicious samples.

The current approach allows including different weighting functions and we discuss several alternatives.

General performance of the novel approach probably will improve as more genomes become available.