Transposable Elements Investigation Tool Applied to Prokaryotic Genomes

Huda Al-Nayyef[1,2], Christophe Guyeux[1], and Jacques M. Bahi[1]
[1] FEMTO-ST Institute, UMR 6174 CNRS, DISC Computer Science Department
Université de Franche-Comté, 16, Rue de Gray, 25000 Besançon, France
[2] Computer Science Department, University of Mustansiriyah, Iraq
{huda.al-nayyef, christophe.guyeux, jacques.bahi}@univ-fcomte.fr

November 2014

Transposable elements (TEs), which are DNA segments that have the ability to insert or copy themselves into new chromosomal locations. In bacterial reign, only cut-and-paste mechanism of transposition can be found, These types of mobile genetic elements (MGEs) involved in such a move being the insertion sequences (ISs). Two main factors have big effects on IS discovery, namely: genes annotation and functionality prediction. The authors have designed a novel pipeline for ISs detection, which embeds the most recently tools, namely OASIS (Optimized Annotation System for Insertion Sequence) [1] and ISFinder database (an up-to-date repository of known ISs) [2].

OASIS identifies ISs in each genome by finding conserved regions surrounding already-annotated transposase genes. It takes as input NCBI genbank file with descriptive functionality. The main problem found in it solved in our pipeline by designing two modules based on OASIS which is called NOASIS and DOASIS. Our pipeline could be represented in the following steps:

Step 1: ORF identification. Our pipeline is currently compatible with any type of annotation tools, having either functionality capability or not, but for comparison we focus on (*BASys*, *Prokka*, and *Prodigal*).

Step 2: IS Prediction. Using either NOASIS or DOASIS for predicting IS elements. Notice that NOASIS requires information about gene functionality by depending not only on NCBI, while DOASIS works with or without gene functionality by modifying genbank files using the suggested methods:

1. **All-Tpase**: we consider that all the genes may potentially be a transposase. So all product fields are set to "transposase".
2. **Zigzag Odd**: we suggest that genes in odd positions are putative transposases and we update the genbank file adequately. Oddly, this

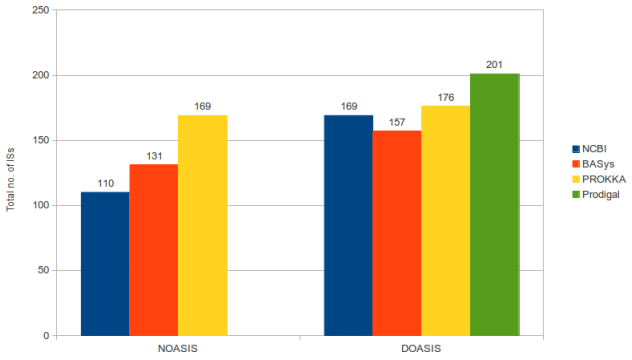

Figure 1: Comparison between NOASIS and DOASIS

new path will produce new candidates which are not detected during All-Tpase.

3. **Zigzag Even**: similar to Zigzag Odd, but on even positions.

Step 3: IS Validation. This step is realized by launching BLASTN on each predicted IS sequence with ISFinder. The e-value of the first hit is then checked: if it is 0.0, then the ORF within this sequence is a Real IS known by ISFinder. It will be considered as Partial IS if its e-value is lower than 10^{-10}. Both IS names of family and group are returned too.

A complete IS detection and classification pipeline has then been proposed and tested on a set of 23 complete genomes of Pseudomonas aeruginosa. This pipeline can also be used as an investigator of annotation tools performance, which has led us to conclude that Prodigal is the suitable annotation tool for IS prediction of prokaryotic. A deepen study regarding IS elements in *P.aeruginosa* has then been conducted, leading to the conclusion that close genomes inside this species have also a close numbers of IS families and groups.

References

[1] David G Robinson, Ming-Chun Lee, and Christopher J Marx. Oasis: an automated program for global investigation of bacterial and archaeal insertion sequences. *Nucleic acids research*, 40(22):e174–e174, 2012.

[2] Patricia Siguier, Jocelyne Pérochon, L Lestrade, Jacques Mahillon, and Michael Chandler. Isfinder: the reference centre for bacterial insertion sequences. *Nucleic acids research*, 34(suppl 1):D32–D36, 2006.