Software Assessment for Prediction of Gene Clusters: An Analysis in silico with Cyanobacteria of Chroococcales Order

Danielle C. C. Couto, Vanessa C. Rezende, Alex R. J. Lima, Felipe C. Couto,

Leonardo T. Dall'Agnol, Evonnildo C. Gonçalves

Federal University of Pará, Federal University of South and Southeast of Pará, State

University of Pará, Federal University of Maranhão

The main objective of this study was the comparison of four different cluster gene prediction tools available (antiSMASH; NP.searcher; NaPDoS; DoBISCUIT) and the influence of the input of the biological information (fasta; annotated; etc). Our work compared the tools using three cyanobacterial genomes from Chrooccocales order: *Cyanobium* CACIAM 14, *Synechocystis* PCC6803 and *Synechocystis* CACIAM 05. The results showed that the integration of the generated data between the different prediction tools promotes deeper and better prospection of clusters. It is important to highlight that depending on the data input format directly influences the number of groups detected, helping to unravel the biotechnological potential of the organisms. The online antiSMASH 3.0.2 in *Cyanobium* sp. CACIAM 14 varied the number of predicted clusters according to the imput: there were 24 clusters for RAST entry against 35 for fasta. Moreover, it is noted that the online version can predict more clusters, due to the incorporation of saccharides and fatty acids clusters. The genome of *Cyanobium* sp. CACIAM 14 was also run on antiSMASH local version 2.0 using gbk and fasta files, annotated with the NCBI PGAP and RAST. Both have generated a total of 15 clusters for gbk and 14 for fasta. Online antiSMASH analysis of *Synechocystis* sp. CACIAM 05 generated as a result 23 clusters for fasta and 16 for gbk. antiSMASH online screening of *Synechocystis* sp. PCC 6803 revealed 33 cluster for fasta and 11 for gbk. However, the result may vary greatly depending on the input type and the types of software used for the genome annotation. The results of NP.searcher tool were more limited, as these cyanobacteria have few (rarely) NRPS/PKS modules. The NaPDoS also had limited results, but has interesting features such as the fact of presenting a tree with the expected product structure, the BLAST results and candidate domains. The DoBISCUIT tool database is extremely extensive and so has a great biotechnological potential related to the prospection of bioactive products in cyanobacteria. Due to this, it presented the highest number of predicted clusters. The mains result of this work was to prove that there are important differences among the results of different gene cluster predictive tools according to the input information and that different tools and parameters should be combined to avoid enlarge the results.