ClustScan and *CompGen* program packages: Semi-automated tools for data mining and homologous recombination modelling

Antonio Starčević, *Faculty of Food Technology and Biotechnology*

Abstract

ClustScan program package is a bioinformatic tool developed mainly for the collection, storage and precise annotation of polyketide, non-ribosomal peptide and other modular biosynthetic gene cluster data hidden in continuously increasing number of sequenced genomes. On the other hand, ***CompGen*** relies on ***ClustScan***'s data in order to simulate homologous recombination events that might happen between gene clusters sequences. We have succeeded to incorporate published and propriety knowledge about modular biosythetic gene clusters in an artificial environment defined by a consistent set of rules. These rules were used for the interpretation of components roles in this environment. The components are the DNA and protein sequences coming from genomes or generated by various programs used by ***ClustScan***. The rules are empirical facts collected from the relevant published scientific papers. As far as we are aware, this system is unique in being able to predict chemical structures from gene cluster DNA sequences and to use known natural products to generate novel "un-natural" ones by homologous recombination simulation. Our future efforts will be to make ***ClustScan*** program package more generic, making it able to scan for gene clusters other than modular ones. Secondly, using ***CompGen***, we will create a database of novel compounds which will allow screening of their biological activity using CADD technologies.