Topic: **Sequence Analysis**

EXPLORING PKS DOMAINS DIVERSITY PRESENT IN ENVIRONMENTAL DATABASE USING HMM PROFILES

R Cuadrat[1], J Cury[2], K Ocaña[1], D Tschoeke[1], A Davila[1]

[1]*1Laboratório de Biologia Computacional e Sistemas - IOC - FIOCRUZ.*
[2]*2Laboratório de Ecologia Microbiana Molecular - UFRJ.*

The overuse of antibiotics in the last decades has caused the emergence of multi-resistant strains (g.e. *Staphylococcus aureus* resistant to vancomycin). Thus, it becomes important the discovery of new molecules with antimicrobial activity. Usually, the search for new molecules is based on microorganisms cultivation. However, only a small fraction (1-10%) of microorganisms in nature can be cultivated ``in vitro", suggesting that the genetic diversity of antimicrobial substances is underestimated. Therefore, the development of new drugs is limited by small diversity of molecules accessed by cultivation methods. Fortunately, in the present days metagenomic methods and bioinformatic tools are available and them can be associated to this aim. The gene families that encode polyketides synthases (PKSs) have biotechnology relevance due to their implication on antimicrobial and antitumor compounds. The aim of this work was to explore the diversity of the conserved PKS domains present in environmental GenBank database. For this, all sequences of three major conserved domains of PKSs (KS - ketoacyl synthase, AT acyl transferase and ACP - acyl carrier protein) were downloaded from the PKSDB (http://linux1.nii.res.in/~pksdb/DBASE/pageALL4.html) and aligned with MAFFT. A total of 60 HMM profiles were built with HMMER (*hmmbuild*) using these multiple alignments. Then, these profiles were calibrated with *hmmcalibrate* and *hmmsearch* was used to search the most similar sequences present in the enviromental database from NCBI. For further analysis there are used the sequences with an e-value smaller than $10e^{-5}$. Sequences that contain at least two PKS domains were parsed using fastacdm (from BLAST package) and the conserved blocks extracted with Gblocks. These conserved blocks were used to construct phylogenetic trees (using MEGA package), with the domains sequences from PKSDB and some FAS I and type II PKS, to distinguish between them. After the HMM search it was possible to find a total of 13465 hits related to KS, of which 9.93% related to MYXALAMIDE synthesis, 8.12% related to EPOTHILONE production and 7.84% related to TYLACTONE production. Of the 12142 hits related to the AT domain, 8.57%, 8.29% and 7.94% are similar to PKS producing of AMPHOTERICIN , STIGMATELLIN and RIFAMYCIN respectively. It was found 848 matches related to ACP, of which 10.14% related to SORAPHEN production, 8.84% related to NIDDAMYCIN production and 8.84% related to EPOTHILONE production. Only 3 sequences with similarity to all domains analyzed were found, all from marine metagenomic (Global Ocean Sampling Expedition). KS, AT and ACP domains are necessary for the existence of functional PKS, then this strongly suggest that sequences presenting all 3 domains may be complete PKSs. BLAST analysis of these sequences against Refseq database showed high similarity hits with PKS (type I and II) in addition to FAS (fatty acid synthase), suggesting that it is not possible to discriminate PKS types and separate it from FAS using only HMM profiles. However, we found 351 hits with only AT and KS, 61 hits with KS and ACP and 22 hits with AT and ACP domains. The abundance of sequences with only 2 domains can be explained due to the origin of sequences, that have low sequencing coverage, producing short environmental sequences that do not contain the three domains. The results show an unexplored diversity of PKS especially in marine environments, suggesting that it is possible to discover new bioactive compounds using metagenomic approaches. Supported by: CNPq/IOC/FIOCRUZ