Abstract

Terpenes and terpenoids represent the largest and most structurally diverse group of natural products, with applications across many fields, including the pharmaceutical industry. These molecules are synthesized in nature by enzymes known as terpene synthases. This thesis conducted a bioinformatic analysis of a curated database containing all 1125 experimentally characterized terpene synthases, focusing on identifying patterns in sequence lengths and domain architectures of these enzymes across different kingdoms of life.

Based on this analysis's knowledge, sequence-guided mining was conducted to identify possible new terpene synthases. Using nearly 5.5 billion protein sequences from various large-scale sequence repositories, the mining resulted in the identification of more than 600 thousand putative terpene synthases. These putative terpene synthases mainly originate from Bacteria and metagenomes, sources that had historically been less explored.

The resulting dataset, accompanied by a phylogenetic tree, sequence similarity network, and two prioritization scores, offers a valuable resource for the discovery of novel terpenes.

Keywords: terpene synthase, TPS, mining, Pfam, SUPERFAMILY, domain, terpene