UNRAVELING THE UNKNOWN UNKNOWNS IN THE METAGENOMIC PROTEIN UNIVERSE USING GRAPHICAL MODELS

Antonio Fernàndez-Guerra[1], Renzo Kottmann[1], Albert Barberán Torrents[2], Frank Oliver Glöckner[1,3] and Emilio O. Casamayor[2]

[1] Microbial Genomics and Bioinformatics Research Group, Max Planck Institute for marine Microbiology, Celsiusstr. 1, 28359 Bremen , Germany
E-mail: rkottman@mpi-bremen.de

[2] Biogeodynamics & Biodiversity Group, Centre d'Estudis Avançats de Blanes, CEAB–CSIC, Accés Cala St. Francesc 14, 17300 Blanes, Spain

[3] Jacobs University, Campus Ring 1, 28759 Bremen, Germany

Metagenomic surveys, like the Global Ocean Survey (GOS), generated a huge amount of genetic data and allow performing more holistic approaches to study marine ecosystems. Moreover, metagenomics proofed being valuable in discovering missing pieces in marine biological processes. However, metagenomics not only expanded our limited view on the diversity of the known protein universe, it also increased the number of genes of unknown functions. Metagenomics reveals a large number of known unknowns like the domains of unknown function (DUF) and unknown unknowns, putative coding sequences without any hint of potential function. Here we propose a novel approach to extract valuable information from the co-occurrence of individual protein domains involved in biological processes in metagenomic complex systems using Graphical Models. Using an integrative approach, we combine the knowledge of the known protein domain families and 16S rDNA with the unknown unknowns to explore the GOS metagenome. As a result, we are able to reveal new associations in biological processes within known protein families and between known protein families and unknowns.

In conclusion, our approach provides a better understanding of the known biological processes and generates a list of candidates from the unknowns or unknown unknowns related with known processes for experimental verification. In some cases we could even suggest specific cultured organisms for performing lab experimental on genes of unknown functions. Thus, our approach might play an important role in bioinformatics biodiscovery pipelines for biotechnology.