Omics dark matter

By L. Bittner and E. Pelletier with the support

15:00-15:10 *Introduction / state of the arts*

15:10-15:40 Invited speaker - Antonio Fernandez-Guerra, *Mining the microbiome: from the known to the unknown*
15:40-16:55 Romain Lannes, *Iterative Safe Homologs Finder*

15:55-16:10 Mohammed-Amin Madoui, *Introducing Metavariant Species for Reference-free and Metagenomic-based Population Genomic Analysis*

16:10-16:25 Nils Giordano, *Co-activity networks reveal the structure of planktonic symbioses in the global ocean*
16:30-17:00 *Coffee break*

17:00-17:30 Invited speaker - Isabelle Callebaut, *Exploring the dark proteome using structural signatures*

17:30-17:45 Stefani DRITSA, *Prediction of candidate disease genes through deep learning on multiplex biological networks*

17:45-18:00 *Conclusion*

Abstracts

Mining the microbiome: from the known to the unknown

A. Fernandez-Guerra[1, 2], C. Vanni[1, 2], M. S. Schechter[1], P. L. Buttigieg [3], M. A. Eren[4, 5], A. Barberan[6] and F. O. Glöckner[1, 2]
[1] Microbial Genomics and Bioinformatics Research Group, Max Planck Institute for Marine Microbiology, Bremen, Germany
[2] Jacobs University Bremen gGmbH, Bremen, Germany
[3] HGF-MPG Group for Deep Sea Ecology and Technology, Alfred Wegener Institute, Helmholtz Centre for Polar and Marine Research, Bremerhaven, Germany
[4] Department of Medicine, University of Chicago, Chicago, IL, United States of America
[5] Josephine Bay Paul Center, Marine Biological Laboratory, Woods Hole, MA, United States of America
[6] Department of Soil, Water, and Environmental Science, University of Arizona, Tucson, AZ, USA

Large-scale metagenomic surveys have generated terabytes of sequence data from a multitude of ecosystems, redefining microbial protein diversity. However, our functional understanding of microbial communities is strongly constrained by the large fraction of still uncharacterized genes. The proportion of this unannotated fraction is considerably large in both sequenced genomes, where it is estimated around 30%, and metagenomes, where it ranges from 40% to the 60% of the genes in the datasets. Unveiling this functionally uncharacterized space has the potential to accelerate the discovery of new functions and help

widen our knowledge of microbial roles in different ecosystems. Despite the several attempts to categorize them, many unknown genes are still found without any classification. We tackled the uncharacterized functional fraction developing a bioinformatic workflow that combines protein clustering with an in-depth evaluation and categorization of the clusters of unknown genes (CUs).
We applied our approach on a comprehensive dataset that combines ~1,900 metagenomes from the marine and human microbiomes, and ~29,000 genomes from the Genome Taxonomy Database (GTDB). As a result, we compiled a database of high-quality clusters, organized in three main categories based on their functional characterization: the Knowns (Ks), clusters annotated to characterized genes, the Genomic unknowns (GUs), clusters of uncharacterized genes found in sequenced or draft genomes, and the Environmental unknowns (EUs), clusters of completely uncharacterized genes found only in environmental samples. We aggregated the clusters into higher-level functional communities to obtain functional units closer to actual protein families. To understand the occurrence and distribution of the CUs in genomes and environment, first, we analyzed the phylogenetic distribution of our clusters on the bacterial and archaeal genomes from the GTDB, following a similar approach as the one used in Annotree. Besides finding lineage-specific CUs at lower taxonomic levels, we identified lineage-specific CUs on high taxonomic levels (class and family) that can represent functional innovations that will help to get a better understanding of the diversification processes. And second, we explored the environmental occurrence of the CUs, which revealed that a large proportion of the CUs present a narrow distribution across sites, suggesting potential adaptive value and supporting their ecological relevance. The same analysis revealed the existence of a ubiquitous fraction of CUs that have the potential to uncover new phylogenetic markers and essential functions.

##########

Iterative Safe Homologs Finder

Romain LANNES[1] , Philippe LOPEZ[1] and Eric Bapteste[1]

[1] Université, Institut de Systématique, Evolution, Biodiversité (ISYEB), Sorbonne Université, CNRS, Museum National d'Histoire Naturelle, EPHE, Université des Antilles, 7, quai Saint Bernard, 75005 Paris, France

Homology detection, i.e. detection of common ancestry, is a standard method for automatic sequence annotation. Homology between sequences that have diverged a long time ago or are fast evolving is not always detectable by direct alignment. Here, we present Iterative Safe Homologs Finder (ISHF), a Python pipeline, that uses an iterative alignment procedure, using previously detected sequences to recover remote homologs of a gene family from a large data set. We investigate the presence of deep branching prokaryotes in the tree of life. We identify putative ancient gene families using sequence similarity networks. We find remote homologs of these ancient gene families using ISHF in a large metagenomic data set, hinting at a potential hidden microbial diversity in environmental data sets.

##########

Introducing Metavariant Species for Reference-free and Metagenomic-based Population Genomic Analysis

Romuald Laso-Jadart[1], Pierre Peterlongo[2], Christophe Ambroise[3], and Mohammed-Amin Madoui[1]
[1] Genomique Métabolique, Institut François Jacob, CEA, CNRS, Univ Evry, Université Paris-Saclay, Evry, France
[2] Univ Rennes, CNRS, Inria, IRISA - UMR 6074, F-35000, Rennes
[3] Univ Evry, Université Paris-Saclay, IBGBI-Lamme, Evry, France

The availability of large metagenomic data offers new opportunities for population genomic analysis of uncultured organisms, especially for small eukaryotes that represent an important part of the unexplored biosphere. However, a very large majority of non-model species lacks reference genome which remains an issue for population genomics.

We introduce the concept of metavariant species (MVS) that is a representation of organisms only by their nucleic polymorphism. To build MVS, metavariants are detected directly from multisample raw metagenomic reads by DiscoSnp++, a reference-free variant caller. Then, the metavariants corresponding to intraspecies polymorphism are clustered based on the depth of coverage of the variable loci using multiple density-based clustering (dbscan). The metavariant clusters are scored according to their expected depth of coverage distribution in each sample and to their size. The metavariant clusters are selected using a maximum weighted independent set (MWIS) algorithm to produce the MVS. The frequencies of the metavariants are then used to perform population genomic analysis on each MVS. The method was implemented in metaVaR, tested on simulated data and compared to other clustering algorithms. metaVaR was also tested on real data containing organisms with available genome and known genomic differentiation.
The method developed in metaVaR showed that the lack of reference genomes or transcriptomes can be bypassed by representing species only by their variable loci directly from raw metagenomic data and that population genomic analysis of multiple unknown species can be performed simultaneously.

##########

Co-activity networks reveal the structure of planktonic symbioses in the global ocean

Nils Giordano[1] and Samuel Chaffron[1]
[1] Laboratoire des Sciences du Numérique de Nantes (LS2N) – CNRS, Université de Nantes, École Centrale de Nantes, IMT Atlantique, 2 rue de la Houssinière, 44322 Nantes, France.

Marine microbes play crucial ecological and biogeochemical roles on our planet, forming the basis of the marine food web, sustaining Earth's biogeochemical cycles in the oceans, and regulating climate. Limited by the fact that most microbes are difficult to isolate and cultivate in lab-controlled environments, meta-omic studies are instrumental to unravel the laws governing the complexity and diversity of their interactions. Today, the amount of data accumulated by large-scale environmental surveys is considerable and significant efforts have been made towards genome reconstruction from metagenomes. So-called Metagenome-Assembled Genomes (MAGs) improve the taxonomic and functional annotation of sequences by binning them into heritable and metabolically viable units. However, little is known about the biotic interactions structuring marine microbial communities. Here, we propose a trait-based approach to uncover putative biotic interactions between MAGs by directly inferring genomic and growth traits from meta-omics data. Available metatranscriptomic data grant access to the expression of bacterial genomes in their environment, while new methods have emerged to infer bacterial replication rates based on differential coverage in a metagenomic sample. Across samples, these co-expression and co-growth signals can thus be exploited to reveal interactions between specific MAGs and link their activities to the environmental context. In addition, we can use the functional content of these co-active MAGs to predict their potential dependencies, in particular if they deviate from general scaling laws that govern the functional content of genomes from lab-cultivated microbial organisms. Inferring and combining (meta-)genomic traits in a global framework can help to identify consortia of marine microbes and pave the way towards the functional understanding and the metabolic modeling of their interactions.

##########

Exploring the dark proteome using structural signatures

Tristan Bitard-Feildel, Alexis Lamiable, Joseph Rebehmed#, Guilhem Faure, Apolline Bruley, Elodie Duprat, Isabelle Callebaut
Sorbonne Université, Muséum National d'Histoire Naturelle, UMR CNRS 7590, Institut de Minéralogie, de Physique des Matériaux et de Cosmochimie, IMPMC, 75005 Paris, France
Lebanese American University, Department of Computer Science and Mathematics, Beirut, Lebanon

A key issue in understanding biological systems is the dark proteome, that is sequences that could not be assigned to known structures or domains, some of which lacking detectable similarity with any known protein (ORFans). Here, we will illustrate how structural signatures can be used for detecting remote relationships, thereby refining the current repertoire of ORFans, as well as revealing functional novelties. Moreover, these structural signatures can also be considered for a comprehensive annotation of proteomes in terms of foldable regions. Hence, different flavors of order can be distinguished in the dark "foldome", from well-folded domains exhibiting structural signatures typical of classical folds to sequences undergoing disorder to order transitions, including atypical architectures, which may represent structural and functional innovation. Moreover, deciphering these signatures and innovations in the dark metagenomic repertoires will increase our understanding of ecological niches and functioning of ecosystems.

##########

Prediction of candidate disease genes through deep learning on multiplex biological networks

Stefani DRITSA[1,2], Thibaud MARTINEZ [1,3], Weiyi ZHANG[1,3,6], Chloé-Agathe AZENCOTT[3,4,5], Antonio RAUSELL[1,2]

[1] Clinical Bioinformatics Laboratory, Imagine Institute, Imagine Institute, Paris Descartes University, Sorbonne Paris Cité, 75015 Paris, France

[2] INSERM UMR 1163, Institut Imagine, 75015 Paris, France

[3] MINES ParisTech, PSL Research University, CBIO-Centre for Computational Biology, 75006, Paris, France

[4] Institut Curie, PSL Research University, 75005, Paris, France,

[5] INSERM, U900, 75005, Paris, France

[6] Shanghai Jiao Tong University

The causal genes or variants of around 50% of all known Mendelian diseases described to date have still not been identified. Network propagation approaches using biological networks, including the human interactome, regulome, phenome and diseasome, have been successfully contributed to the discovery rate on new disease genes. In this study we propose to leverage recent deep learning advances in semi-supervised node labeling to address multiplex network integration in disease gene discovery.

To that aim we implemented a model that, taking a given number of networks and a node class as input, uses both unsupervised learned embeddings and supervised graph-structure learning. Thus, for a collection of (un)directed and (un)weighted graphs G=(V,E,R), with nodes vi Î V, edges (vi,r,vi) Î E of relation type r Î R (number of nodes |V|=N, number of relations |R|=R), we perform node representation learning through node2vec and run normalized Relational graph convolutional networks (R-GCNs) on all networks. Algorithmic novelties to incorporate node attributes, attention mechanisms, and unsupervised clustering into the learning process have been developed.

The model uses as input a collection of more than 100 biological networks, including protein-protein interactions, tissue-specific gene regulatory and co-expression networks, signaling networks, and functional similarity networks based on different ontologies. It was hence tested by its ability to prioritize Mendelian diseases genes of different categories and benchmarked against reference state-of-the-art methods. Detailed examples are presented and results interpreted in the context of the associated local network topologies.