Title of the PhD project:

Deep learning to predict metabolic protein-protein interaction networks

for environmental microbial communities

PhD Supervisor

Name, first name : Carbone, Alessandra
Email : alessandra.carbone@lip6.fr
Phone : 01.44.27.73.45
Title/Employer : Sorbonne Université
Address : Campus Pierre & Marie Curie, bât C, 4 place Jussieu, 75005 Paris
Laboratory : Laboratoire de Biologie Computationnelle et Quantitative (LCQB), UMR7238
Title of the team : Génomique Analytique
Team leader (*if different*) *:*
Doctoral School : ED130-EDITE

Overview of the scientific projects of the team

L'équipe « Génomique Analytique » développe des approches mathématiques, issues des statistiques et de la combinatoire, et des outils algorithmiques pour étudier les principes de base du fonctionnement cellulaire à partir de données génomiques. Nos projets visent à comprendre les principes de base de l'évolution et de la co-évolution des structures moléculaires dans la cellule. Ils concernent : séquences et structures protéiques, phénotypes protéiques et mutations génétiques, organisation du génome. Les applications sont multiples et jouent un rôle dans la mutagenèse dirigée, la biologie synthétique, l'organisation des données métagénomiques et l'annotation du génome.

Main publications since January 1er, 2016

1. G. Drillon, R. Champeimont, F. Oteri, G. Fischer, **A. Carbone**, "Phylogenetic reconstruction based on synteny block and gene adjacencies". *Molecular Biology and Evolution*, 2020.

2. E. Laine, Y. Karami, **A. Carbone**, "GEMME: a simple and fast global epistatic model predicting mutational effect", *Molecular Biology and Evolution*, 2019.

3. F. Douam, F. Fusil, M. Enguehard, L. Dib, F. Nadalin, L. Schwaller, J. Mancip, L. Mailly, T. Baumert, **A. Carbone***, FL. Cosset*, D. Lavillette*. "A protein coevolution method designed for conserved sequences uncovers critical features of the original HCV fusion mechanism and

provides molecular basis for the design of effective antiviral strategies". *PLoS Pathogens*, 2018. (* corresponding author)

4. A. Ugarte, R. Vicedomini, J. Bernardes, **A. Carbone**, MetaCLADE: a multi-source annotation method for metagenomic and metatranscriptomic sequences. *Microbiome*, 2018.

5. F. Oteri, F. Nadalin, R. Champeimont, **A. Carbone**, "BIS2Analyzer: a server for coevolution analysis of conserved protein families". *Nucleic Acids Research*, 2017.

PhD Co-Supervisor

Name, first name : Bittner, Lucie
Email : lucie.bittner@upmc.fr
Phone : 01 40 79 48 07
Title/Employer : Sorbonne Université
Address : Muséum National d'Histoire Naturelle, Bâtiment d'Entomologie, 45 rue Buffon, 75005 Paris
Laboratory : Institut de Systématique, Evolution, Biodiversité - ISYEB UMR7205
Title of the team : Atelier de BioInformatique
Team leader *(if different) :* co-direction Mathilde Carpentier et Lucie Bittner
Doctoral School : ED227 Sciences de la nature et de l'Homme : évolution et écologie

Overview of the scientific projects of the team

L'équipe "Atelier de BioInformatique" (ABI) est une structure ouverte regroupant des enseignants-chercheurs et chercheurs travaillant à l'interface entre Biologie et BioInformatique dans un environnement multidisciplinaire. L'ABI est fortement investi dans la formation universitaire (enseignements en bioinformatique et biologie à Sorbonne Université et formation à la recherche de masters, doctorants et post-doctorants) et accueille régulièrement toute personne souhaitant traiter son sujet de recherche par l'approche bioinformatique. Les thèmes principaux de recherche pour lesquels des outils bioinformatiques sont développés couvrent l'évolution moléculaire, la génomique comparée, la structure et l'évolution des protéines, et la génomique environnementale. L Bittner a rejoint l'ABI en janvier 2019, ce qui a permis de renforcer significativement les dynamiques de recherche en génomique environnementale et évolutive.

Main publications since January 1er, 2016

Faure E., Not F., Benoiston A.-S., Labadie K., **Bittner L**.*, Ayata S.-D.* (co-last authors) (2019). Mixotrophic protists display contrasted biogeographies in the global ocean. *The ISME Journal* **13**(4):1072-1083. https://doi.org/10.1038/s41396-018-0340-5

Guidi L.*, Chaffron S.*, **Bittner L.***, Eveillard D. (* co-first authors), Larhlimi A., Roux S., Darzi Y., Audic S., Berline L., Brum J.R., Coelho L.P., Espinoza J.C.I., Malviya S., Sunagawa S., Dimier C., Kandels-Lewis S., Picheral M., Poulain J., Searson S., Tara Oceans Consortium Coordinators, Stemmann L., Not F., Hingamp P., Speich S., Follows M., Karp-Boss L., Boss E., Ogata H., Pesant S., Weissenbach J., Wincker P., Acinas S.G., Bork P., de Vargas C., Iudicone D., Sullivan M.B., Raes J., Karsenti E., Bowler C., Gorsky G. (2016). Plankton networks driving carbon export in the oligotrophic ocean. *Nature* **532**, 465–470. https://doi.org/10.1038/nature16942

Lewitus E., **Bittner L**., Malviya S., Bowler C., Morlon, H. (2018). Clade-specific diversification dynamics of marine diatoms since the Jurassic. *Nature Ecology & Evolution* **2**, 1715–1723. https://doi.org/10.1038/s41559-018-0691-3

Meng A., Corre E., Probert I., Gutierrez-Rodriguez A., Siano R., Annamale A., Alberti A., Da Silva C., Wincker P., Le Crom S., Not F., **Bittner, L.** (2018a). Analysis of the genomic basis of functional diversity in dinoflagellates using a transcriptome-based sequence similarity network. *Molecular Ecology* **27**(10): 2365-2380. https://doi.org/10.1111/mec.14579

Meng A., Marchet C., Corre E., Peterlongo P., Alberti A., Da Silva C., Wincker P., Pelletier E., Probert I., Decelle J., Le Crom S., Not F., **Bittner L.**, 2018b. A de novo approach to disentangle partner identity and function in holobiont systems. *Microbiome* **6**(1), 105. https://doi.org/10.1186/s40168-018-0481-9

Doctoral Project

Title: Deep learning to predict metabolic protein-protein interaction networks for environmental microbial communities

Abstract: Metagenomics provides a huge inventory of species present in environments and of metabolic functions performed by environmental communities. It offers enormous potential for discoveries, as more than 99% of microbial species cannot be cultivated in the laboratory. It has given rise to several large-scale projects to characterize the microbial diversity of the oceans (e.g., GOS [1], Tara Oceans [2], Malaspina [3], OSD [4]), of microbes in symbiosis with humans [5], of the composition of soils [6], urban environments [7], or subjected to extreme conditions (eXtreme Microbiome Project). Each metagenomics experiment generates large amounts of raw data (on the order of several terabytes of sequence per sample), the processing of which presents several algorithmic and data analysis / learning challenges.

The goal of this PhD project will be to **develop new computational approaches** based on **deep learning** to reconstruct **protein-protein interaction (PPI) networks** for **metagenomic samples starting from sequence reads. The aim is to predict PPI networks that** allow a **community of microbes** to perform their metabolic functions. Questions on biogeography and evolution of PPIs will be addressed with a comparison of PPI though samples from different ecosystems.

Context and objective:

Context.

Describing functions in environmental communities is of paramount importance for understanding the balance of the intestinal flora, bioremediation, the mechanisms of resistance to antibiotics, the development of energy-producing bacteria. In this thesis, we want to **characterize environmental metabolic functions through the analysis of the interactions of the genes** present in a community. This requires 1. the annotation of protein domains in sequences, 2. the functional characterizations of new homologous sequences and 3. the inference of protein interactions. These tasks are difficult and complicated by the low sensitivity of current methods.

In the last years, AC team worked on the first two problems and proposed three novel computational approaches. Two are dedicated to a precise domain annotation of genomes (CLADE [8]) and metagenomes (MetaCLADE [9]) and are based on large libraries of probabilistic models. The third is dedicated to functional classification of homologs (ProfileView [10]). These studies provide an optimal starting point to address the third question, which is the reconstruction of the protein-protein interactions for metagenomic sequences and its use in interpreting the behavior of proteins in environmental communities.

Objectives:

1. **We shall design a Deep Learning (DL) architecture especially devoted to handle metagenomic sequences, that given a pair of protein sequences will say whether they are likely partners or not**. This architecture will be based on a DL approach for the inference of protein-protein interaction networks that is under development in the Analytical Genomic lab for sets of proteins for which known small linear interaction motifs might be available. Figure 1 illustrates the exceptional results we obtained on the reference Mintseris dataset [11] of 168 proteins. The performance overpasses existing DL attempts [12] as well as alternative approaches based on complete-cross docking [13]. Not least, it scales to thousands of proteins. The new architecture that will be developed in this thesis will be enriched with knowledge on general properties of biological metagenomic data, on multiple probabilistic models associated to protein domains found in metagenomic samples, on how proteins interact "socially", that is, if they like to interact or avoid interactions (see dark and clear columns in Fig 1). The architecture will be constituted by two neural networks, each dedicated to learning a protein in the pair. The convolution module of the architecture will be designed as a set of 100-200 small filters either randomly defined or encoding randomly chosen small linear motifs coming from an analysis of conserved small linear motifs extracted from metagenomic sequences of domains of interest realized in this project. We shall also add biological information considering residues known or expected to belong to protein binding sites. Note that the first layer composed of many small filters (similar to [14]) helps to learn specific motifs describing protein interactions.

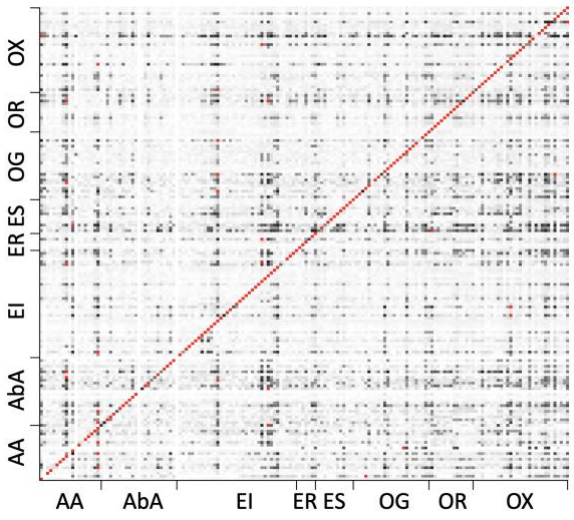

Figure 1 Matrix reporting interaction scores of our recent DL architecture (dark vs light grey indicate strong vs weak interaction; the highest in red) for 168 proteins in the Mintseris dataset. Partners (diagonal) are organised in functional classes (antibody-(bound-) antigen, enzyme with inhibitor/ regulatory chain/ substrate, G-protein / receptor /others containing complexes). The matrix describes an AUC of 94% and a PRAUC of 93%.

2. **Construction of a database of small linear motifs that appear in metagenomic sequences**. It will provide metagenomic information to be included in the DL architecture. Its construction will be carried out starting from a set of several probabilistic models [8,9] constructed from metagenomic sequences, specific to protein domains identified in multiple environments. The database will provide a reference for exploring and classifying protein sequences in metagenomic samples, and will help the prediction of binding sites. Interaction motifs will be used to define the starting filters in the DL architecture. The database will be made available to the community.

3. **Screening of the results from the DL architecture (in 1) towards the construction of an interaction network.** The results obtained with our DL architecture in 1 will be crossed with other biological information to augment confidence in PPI prediction:

- the expected domain co-occurrence in proteins;

- physico-chemical properties of amino acids and genetic diversity observed within metagenomic samples.

In the same way that phylogenetic profiling methods [15] have made it possible to segregate groups of homologous genes associated with the same phenotypes, we will use information on protein domain co-occurrence and combine it with genetic information from correlated mutations to determine which proteins are in direct interaction. Our working hypothesis is that two proteins whose domains are in interaction should: (i) have correlated abundances between several environmental conditions and (ii) exhibit compensatory mutations at their interface across different species and strains that possess them.

(i) MetaCLADE [9] will be used to annotate metagenomic datasets with all domains of the Pfam database. The S3A tool [16] will be applied to perform a targeted assembly of metagenomes, centered around protein domains. Covariation of coverages across the environmental conditions for domains of different proteins will indicate a functional relationship between them. For this task, we will use our recent approach to functional classification ProfileView [10].

(ii) The study of compensatory mutations (eg coevolution) allows to identify contact points between amino acids in protein sequences [17,18,19,20]. We intend to use this principle and develop a second neural network architecture that takes into account the characteristics of the coevolution signals that could be deduced from metagenomic sequence profiles [21]. By combining information on domain co-occurrence [22,8] with the presence of compensatory mutations on profiles, we will provide further confidence in the protein-protein interaction network. This analysis will be achieved at a domain level, thus providing information on protein domain interaction.

4. **Biogeography and Evolution of the PPI**. From the previous steps, a list of PPI and their presence/abundance into environments will be produced. Based on these data and information related to sampled ecosystems (e.g., sampled ecosystems (ocean, soil, gut), temperature, pH, nutrient concentrations, carbon export, net primary production), multivariate statistical analyses will be conducted in order to investigate the biogeography of PPIs [23,24]. Ubiquitous versus endemic PPIs will be highlighted. In particular, we propose to define subsets of PPI, which will be best predictors of given environmental niches, and therefore, which could be used, beyond this project, as inputs for metabolic reconstruction. Furthermore, the exhaustive list of PPIs resulting from step 3 could be analyzed in a phylogenetic framework, which will offer a new and original path to study associations within ecosystems. For example, nodes of the PPI networks will be tagged with taxonomic information (i.e. Bacteria, Archaea, Virus, Eukaryotes, or lower rank), which will allow to quantify the proportion of intra and interphyla associations. Known symbioses will be sought, as well as new ones could be suggested at the ecosystem level (e.g. [25]). By mapping these associations on phylogenetic trees, we will infer when these associations appeared and quantify the variation in the rate of appearance over time.

Justification of suitability for *i-Bio*:
The complementarity of the two advisors, one coming from computer science and the other from biology, will allow for an optimal realization of this project that requires both a competence in the development of Deep Learning architectures for dealing with large quantities of metagenomic data and a competence in the biological analysis of these data to infer meaningful biological results from them.

Role of each supervisor / skills provided:

AC will advise the student on the computation part for the development of deep learning architectures and the statistical analysis of metagenomic data. LB will advise the student on the biological interpretation of her/his results, and guide her/him with the biogeographical and evolutionary studies. We intend to have regular weekly meetings between the three of us to follow the progress of the thesis.

Profile of the desired student:

Computer science or mathematics. This profile is required by the processing of metagenomic data, the development of deep learning architectures and, possibly, of efficient methods (algorithms on strings, compression,…) dedicated to GPU architectures for neural networks.

References

1. Rusch, D.B., Halpern, A.L., Sutton, G., Heidelberg, K.B., Williamson, S., Yooseph, S., Wu, D., Eisen, J.A., Hoffman, J.M., Remington, K., Beeson, K., Tran, B., Smith, H., Baden-Tillson, H., Stewart, C., Thorpe, J., Freeman, J., Andrews-Pfannkoch, C., Venter, J.E., Li, K., Kravitz, S., Heidelberg, J.F., Utterback, T., Rogers, Y.-H., Falcón, L.I., Souza, V., Bonilla-Rosso, G., Eguiarte, L.E., Karl, D.M., Sathyendranath, S., Platt, T., Bermingham, E., Gallardo, V., Tamayo-Castillo, G., Ferrari, M.R., Strausberg, R.L., Nealson, K., Friedman, R., Frazier, M., Venter, J.C., 2007. The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific. *PLOS Biology* 5, e77. https://doi.org/10.1371/journal.pbio.0050077
2. Karsenti, E., Acinas, S. G., Bork, P., Bowler, C., De Vargas, C., Raes, J., Sullivan, M., Arendt, D., Benzoni, F., Claverie, J.-M., Follows, M., Gorsky, G., Hingamp, P., Iudicone, D., Jaillon, O., Kandels-Lewis, S., Krzic, U., Not, F., Ogata, H., Pesant, S., Reynaud, E. G., Sardet, C., Sieracki, M. E., Speich, S., Velayoudon, D., Weissenbach, J., Wincker, P., and the Tara Oceans Consor- tium (2011). A Holistic Approach to Marine Eco-Systems Biology. *PLoS Biology*, 9(10):e1001177.
3. Duarte, C. M. (2015). Seafaring in the 21st century: the Malaspina 2010 circumnavigation expe-dition. *Limnology and Oceanography Bulletin*, 24(1):11–14
4. Kopf, A., Bicak, M., Kottmann, R., Schnetzer, J., Kostadinov, I., Lehmann, K., Fernandez-Guerra, A., Jeanthon, C., Rahav, E., Ullrich, M., Wichels, A., Gerdts, G., Polymenakou, P., Kotoulas, G., Siam, R., Abdallah, R.Z., Sonnenschein, E.C., Cariou, T., O'Gara, F., Jackson, S., Orlic, S., Steinke, M., Busch, J., Duarte, B., Caçador, I., Canning-Clode, J., Bobrova, O., Marteinsson, V., Reynisson, E., Loureiro, C.M., Luna, G.M., Quero, G.M., Löscher, C.R., Kremp, A., DeLorenzo, M.E., Øvreås, L., Tolman, J., LaRoche, J., Penna, A., Frischer, M., Davis, T., Katherine, B., Meyer, C.P., Ramos, S., Magalhães, C., Jude-Lemeilleur, F., Aguirre-Macedo, M.L., Wang, S., Poulton, N., Jones, S., Collin, R., Fuhrman, J.A., Conan, P., Alonso, C., Stambler, N., Goodwin, K., Yakimov, M.M., Baltar, F., Bodrossy, L., Van De Kamp, J., Frampton, D.M., Ostrowski, M., Van Ruth, P., Malthouse, P., Claus, S., Deneudt, K., Mortelmans, J., Pitois, S., Wallom, D., Salter, I., Costa, R., Schroeder, D.C., Kandil, M.M., Amaral, V., Biancalana, F., Santana, R., Pedrotti, M.L., Yoshida, T., Ogata, H., Ingleton, T., Munnik, K., Rodriguez-Ezpeleta, N., Berteaux-Lecellier, V., Wecker, P., Cancio, I., Vaulot, D., Bienhold, C., Ghazal, H., Chaouni, B., Essayeh, S., Ettamimi, S., Zaid, E.H., Boukhatem, N., Bouali, A., Chahboune, R., Barrijal, S., Timinouni, M., El Otmani, F., Bennani, M., Mea, M., Todorova, N., Karamfilov, V., ten Hoopen, P., Cochrane, G., L'Haridon, S., Bizsel, K.C., Vezzi, A., Lauro, F.M., Martin, P., Jensen, R.M., Hinks, J., Gebbels, S., Rosselli, R., De Pascale, F., Schiavon, R., dos Santos, A., Villar, E., Pesant, S., Cataletto, B., Malfatti, F., Edirisinghe, R., Silveira, J.A.H., Barbier, M., Turk, V., Tinta, T., Fuller, W.J., Salihoglu, I., Serakinci, N., Ergoren, M.C., Bresnan, E., Iriberri, J., Nyhus, P.A.F., Bente, E., Karlsen, H.E.,

Golyshin, P.N., Gasol, J.M., Moncheva, S., Dzhembekova, N., Johnson, Z., Sinigalliano, C.D., Gidley, M.L., Zingone, A., Danovaro, R., Tsiamis, G., Clark, M.S., Costa, A.C., El Bour, M., Martins, A.M., Collins, R.E., Ducluzeau, A.-L., Martinez, J., Costello, M.J., Amaral-Zettler, L.A., Gilbert, J.A., Davies, N., Field, D., Glöckner, F.O., 2015. The ocean sampling day consortium. *Gigascience* 4. https://doi.org/10.1186/s13742-015-0066-5

5. The Human Microbiome Project Consortium, Structure, function and diversity of the healthy human microbiome (2012), *Nature* 486, 207–214
6. Fierer et al. Cross-biome metagenomic analyses of soil microbial communities and their functional attributes. (2012), *Proc Natl Acad Sci.* 109(52):21390-5.
7. MetaSUB International Consortium. The Metagenomics and Metadesign of the Subways and Urban Biomes (MetaSUB) International Consortium inaugural meeting report. (2016) *Microbiome.* Jun 3;4(1):24
8. Bernardes J., Zaverucha G., Vaquero C. and Carbone A. (2016) Improvement in Protein Domain Identification Is Reached by Breaking Consensus, with the Agreement of Many Profiles and Domain Co-occurrence. *PLoS Comput Biol*. 12(7):e1005038.
9. Ugarte A., Vicedomini R., Bernardes J. and Carbone A. (2018) MetaCLADE: a multi-source annotation method for metagenomic and metatranscriptomic sequences. *Microbiome*.
10. Vicedomini R., Bouly J.P., Laine E., Falciatore A. and Carbone A. (2019) ProfileView: multiple probabilistic models resolve protein families functional diversity. [**bioRxiv**]
11. Mintseris J., Wiehe K., Pierce B., Anderson R., Chen R., Janin J. and Weng Z. (2005) Protein-Protein Docking Benchmark 2.0: an update. *Proteins* 60:214–216.
12. Hashemifar S., Neyshabur B., Khan A. A. and Xu, J. (2018) Predicting protein–protein interactions through sequence-based deep learning. *Bioinformatics*, *34*(17), i802-i810.
13. Lopes A., Sacquin-Mora S., Dimitrova V., Laine E., Ponty Y. and Carbone A. (2013) Protein-protein interactions in a crowded environment: an analysis via cross-docking simulations and evolutionary information, *PLoS Computational Biology*.
14. Kelley D. R., Snoek J. and Rinn J. L. (2016) Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome research*, *26*(7), 990-999.
15. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. (1999) *Proc Natl Acad Sci*, Apr 13;96(8):4285-8.
16. David L., Vicedomini R., Richard H. and Carbone A. (2020) Targeted domain assembly for fast functional profiling of metagenomic datasets with S3A. *Bioinformatics*.
17. Lockless S.W. and Ranganathan R. (1999) Evolutionarily Conserved Pathways of Energetic Connectivity in Protein Families. *Science* 286: 295-299.
18. Weigt M., White R.A., Szurmant H., Hoch J.A. and Hwa T. (2009) Identification of direct residue contacts in protein-protein interaction by message passing. *Proc. Natl. Acad. Sci.* 106, 67.
19. Carbone A. and Dib L. (2010) Co-evolution and information signals in biological sequences. *Theoretical Computer Science*. doi:10.1016/j.tcs.2010.10.040.
20. Oteri F., Nadalin F., Champeimont R. and Carbone A. (2017) BIS2Analyzer: a server for coevolution analysis of conserved protein families. *Nucleic Acids Research*.
21. Eddy S.R. (1998) Profile hidden Markov models. *Bioinformatics*. 14(9):755-63.
22. Bernardes J., Vieira F.R.J., Zaverucha G. and Carbone A. (2015) A multi-objective optimisation approach accurately resolves protein domain architectures. *Bioinformatics*.

23. Faure E., Not F., Benoiston A.-S., Labadie K., Bittner L.*, Ayata S.-D.* (co-last authors) (2019). Mixotrophic protists display contrasted biogeographies in the global ocean. *The ISME Journal* 13(4):1072-1083. https://doi.org/10.1038/s41396-018-0340-5

24. Faure E., Ayata S.-D., Bittner L., Towards omics-based predictions of planktonic functional composition from environmental data. *In revision* for *Nature Communications*.

25. Lima-Mendez, G., Faust, K., Henry, N., Decelle, J., Colin, S., Carcillo, F., Chaffron, S., Ignacio-Espinosa, J.C., Roux, S., Vincent, F., Bittner, L., Darzi, Y., Wang, J., Audic, S., Berline, L., Bontempi, G., Cabello, A.M., Coppola, L., Cornejo-Castillo, F.M., d'Ovidio, F., Meester, L.D., Ferrera, I., Garet-Delmas, M.-J., Guidi, L., Lara, E., Pesant, S., Royo-Llonch, M., Salazar, G., Sánchez, P., Sebastian, M., Souffreau, C., Dimier, C., Picheral, M., Searson, S., Kandels-Lewis, S., Coordinators, T.O., Gorsky, G., Not, F., Ogata, H., Speich, S., Stemmann, L., Weissenbach, J., Wincker, P., Acinas, S.G., Sunagawa, S., Bork, P., Sullivan, M.B., Karsenti, E., Bowler, C., Vargas, C. de, Raes, J., 2015. Determinants of community structure in the global plankton interactome. *Science* 348, 1262073. https://doi.org/10.1126/science.1262073