Journée réseau 2025

22 mai 2025 MIO, Campus de Luminy

Marseille, France

réseau BIOinformaTIque en provenCe

Préface

Cher·e·s participant·e·s, cher·e·s collègues,

Bienvenue à la journée du réseau BIOinformaTIque en provenCe (BIOTIC) ! Nous sommes très heureux de vous accueillir pour cette édition, qui affiche complet avec 80 participant·e·s. Plusieurs d'entre vous contribuent activement à cette journée, que ce soit à travers une communication orale ou en assurant la présidence d'une session, et nous vous en remercions vivement.

Nous avons également l'honneur d'accueillir trois invités spéciaux, Céline Brochier, Étienne Danchin et William Ritchie, tous d'anciens Marseillais, dont les expertises dans différents domaines de la bioinformatique enrichiront les échanges tout au long de la journée.

Nous espérons que vous apprécierez la variété des thématiques abordées et que cette rencontre sera l'occasion de renforcer les liens au sein de notre communauté scientifique régionale.

Très bonne journée à toutes et à tous !

Benoît Ballester (TAGC)
Delphine Potier (CRCM)
Fabrice Armougom (MIO)
Romain Fenouil (CIML)
Vincent Lombard (AFMB)

Table des matières

Programme

Horaires	Évènement	Chair	Lieu
08:30 - 09:00	Accueil		
09:00 - 09:10	Présentation de la journée, les nouvelles du réseau - CoPil BIOTIC		
09:10 - 10:00	Identification of thermoadaptation fingerprints in three-dimensional protein structures using machine learning and persistent homology *Céline BROCHIER-ARMANET (LBBE)*	Gaël Erauso (MIO)	
10:00 - 10:20	Tracking down microbial primary producers from the serpentinite-hosted Prony Bay hydrothermal field via metagenomics *Rabja Popall (MIO)*		
10:20 - 10:50	Pause café		
10:50 - 11:40	Sécurité des données, parcimonie et spécificité. De nouveaux outils d'IA pour lutter contre les problèmes persistants. *William RITCHIE (IGH)*	Matthieu Legendre (IGS)	
11:40 - 12:00	Benchmarking Data Leakage on Link Prediction in Biomedical Knowledge Graph Embeddings *Galadriel BRIERE (MMG)*		
12:00 - 12:25	Flash talks *G. Bidaut, M. Boulinguiez, M. Garel, F. Ghestem, P. Kousteridou, B. Loire, L. Quatreville, M. Slivak*		
12:25 - 14:00	Déjeuner & Session poster		
14:00 - 14:50	End of the beginning: long-read genome assemblies of hybrid parasitic worms reveal unusual chromosome ends *Etienne DANCHIN (Sophia Agrobiotech)*	Pedro Coutinho (AFMB)	
14:50 - 15:10	Exonic enhancers are a widespread class of dual-function regulatory elements *Jean-Christophe Mouren (TAGC)*		
15:10 - 15:30	Discussion autour de la bioinformatique pour le GT stratégie scientifique AMU *Christine Brun (TAGC)*		
15:30 - 16:00	Pause café		
16:00 - 16:20	Dissecting microbe-host relationships via interface-resolved protein interaction networks *Andreas Zanzoni (TAGC)*	Emmanuel Talla (LCB)	
16:20 - 16:40	Detect de novo expressed ORFs in transcriptomes with DESwoMAN *Anna Grandchamp (TAGC)*		
16:40 - 17:00	Cloture - CoPil BIOTIC		

Conférencier.e.s invité.e.s

Céline Brochier

Laboratoire de Biométrie et Biologie Evolutive (LBBE), Lyon

"Identification of thermoadaptation fingerprints in three-dimensional protein structures using machine learning and persistent homology"

Etienne Danchin

Genomics & Adaptive Molecular Evolution (GAME) à l'Institut Sophia Agrobiotech, Sophia-Antipolis

"End of the beginning: long-read genome assemblies of hybrid parasitic worms reveal unusual chromosome ends"

William Ritchie

Institut de Génétique Humaine (IGH), Montpellier

"Sécurité des données, parcimonie et spécificité. De nouveaux outils d'IA pour lutter contre des problèmes persistants"

Identification of thermoadaptation fingerprints in three-dimensional protein structures using machine learning and persistent homology

Léa Bou Dagher[1,2,3], Gabriela Ciuperca[2], Dominique Madern[4], Philippe Oger[5], Priscilla Gillet[5], Philippe Malbos[2], Céline Brochier-Armanet[*1]

[1] Université Claude Bernard Lyon 1, CNRS, VetAgro Sup, Laboratoire de Biométrie et Biologie Évolutive, UMR5558, Villeurbanne, France
[2] Université Claude Bernard Lyon 1, CNRS, Institut Camille Jordan, UMR5208, Villeurbanne, France
[3] Université Libanaise, Laboratoire de mathématiques, École Doctorale en Science et Technologie, PO BOX 5, Campus Rafik Hariri, Hadath, Liban
[4] Structural Biology Group, European Synchrotron Radiation Facility, 71 Avenue des Martyrs, 38000 Grenoble, France
[5] Univ Lyon, INSA Lyon, CNRS, UMR5240 Microbiologie Adaptation et Pathogénie, F-69621 Villeurbanne, France

Résumé

Proteins contain a phylogenetic signal that allows the history of organisms to be traced, but they also contain information about the environmental and genomic constraints to which they are subject. Until recently, most evolutionary studies were based on protein sequence analysis (> 250 million sequences available in Uniprot, compared to 230,000 experimentally resolved structures in the PDB). The recent development of reliable methods for predicting three-dimensional protein structures opens new perspectives (> 250 million structures available). However, this avalanche of data requires new analysis methods. In this context, we have initiated a new project at the intersection of topological data analysis, bioinformatics and molecular biology with the aim to develop: (i) new geometric representations of structures (called biogeometric markers) (1), (ii) new methods based on persistent homology to analyse these representations (2, 3), and new prediction models based on machine learning. These developments will be illustrated using the example of thermoadaptation, an adaptive process that strongly constrains the evolution of proteins, in particular the abundancy of certain amino acids in proteins. As a consequence, strains with different optimal growth temperatures have proteomes with different amino acid compositions (4, 5). However, environmental temperature is not the only factor influencing protein amino acid composition, and factors such as genomic GC content, salinity, and expression level also have an effect. In a recent study, we showed that Methanococcales is an interesting model to study thermoadaptation (6, 7). Indeed, in these archaea, the optimal growth temperature is the main factor influencing amino acid abundance in proteomes, explaining 70% of the observed variance, whereas in other prokaryotic lineages it explains at most 20%. By coupling molecular phylogenetics with ancestral

[*] Intervenant

sequence reconstruction, we have unravelled the underlying substitution patterns, revealing lysine as a key player in this process (6). We have also shown that the protein structures do indeed contain a strong signal, which has allowed us to build a molecular thermometer that can predict the optimal growth temperature of strains from individual protein structures with a margin of error of ±5°C and an accuracy greater than 0.90.

1. L. Bou Dagher, D. Madern, P. Malbos, C. Brochier-Armanet, Faithful Interpretation of Protein Structures through Weighted Persistent Homology Improves Evolutionary Distance Estimation. *Mol Biol Evol* 42 (2025).
2. L. Bou Dagher, D. Madern, P. Malbos, C. Brochier-Armanet, Persistent homology reveals strong phylogenetic signal in three-dimensional protein structures. *PNAS Nexus* in press (2024).
3. G. Carlsson, Topology and data. *B Am Math Soc* 46, 255-308 (2009).
4. K. B. Zeldovich, I. N. Berezovsky, E. I. Shakhnovich, Protein and DNA sequence determinants of thermophilic adaptation. *PLoS computational biology* 3, e5 (2007).
5. B. Boussau, S. Blanquart, A. Necsulea, N. Lartillot, M. Gouy, Parallel adaptations to high temperatures in the Archaean eon. *Nature* 456, 942-945 (2008).
6. M. Lecocq, M. Groussin, M. Gouy, C. Brochier-Armanet, The Molecular Determinants of Thermoadaptation: Methanococcales as a Case Study. *Molecular biology and evolution* 38, 1761-1776 (2021).
7. D. Madern et al., The Characterization of Ancient Methanococcales Malate Dehydrogenases Reveals That Strong Thermal Stability Prevents Unfolding Under Intense gamma-Irradiation. *Mol Biol Evol* 41 (2024).

End of the beginning: long-read genome assemblies of hybrid parasitic worms reveal unusual chromosome ends

Etienne GJ Danchin*[1] *et al.*

[1]*Institut Sophia Agrobiotech, INRAE, Université Côte d'Azur, CNRS, 400 route des Chappes, Sophia Antipolis*

Résumé

Telomeres are nucleoprotein complexes that cap linear chromosomes and protect them from fusion and degradation. They are involved in cell ageing regulation and their dysfunction can cause serious disease. In the model nematode *C. elegans* telomeric DNA is made of (TTAGGC)n terminal repeats associated with single-strand as well as double-strand DNA binding proteins, forming a protective terminal complex. Telomeric repeats are added at chromosome ends after DNA replication by a telomerase reverse transcriptase, using an RNA template. This system is assumed to be widely conserved in eukaryotes, including in the phylum Nematoda. Using long-read sequencing, we have assembled the genomes of the three most economically important root-knot nematodes (genus Meloidogyne). *Meloidogyne incognita*, *M. javanica* and *M. arenaria* are devastating plant pests with polyploid (3n – 4n) genomes as a result of complex interspecific hybridizations, which poses challenges for correct separation of the haplotypes. We have assembled the genomes at high contiguity levels, with N50 values of ~2Mb and have mostly unzipped the assembly in A and B sub-genomes. The biggest contigs represent nearly complete chromosomes, allowing investigations of how they start and end. The canonical (TTAGGC)n repeat was not found in any of the *Meloidogyne* genomes analyzed and no evidence for a telomerase or orthologs of *C. elegans* telomere-binding proteins could be found. Instead, bioinformatics analyses revealed complex motifs a*t one end of several contigs. Using DNA FISH experiments, we revealed that these complex motifs were mostly at one end of chromosomes in the three species. These complex repeats are specific to mitotic parthenogenetic root-knot nematodes and return no significant similarity to any other species. Yet, they present several characteristics of *bona fide* telomeric repeats, including the ability to form G-quadruplex, their stranded orientation and evidence for transcription. This ensemble of results suggests mitotic parthenogenetic root-knot nematodes possess very specific complex motifs at one end of their chromosomes. Proteins and RNA molecules interacting with these repeats remain to be discovered. These findings open new perspectives towards understanding how genome integrity is maintained in these polyploid mitotic pests of worldwide agricultural importance.

* Intervenant

Sécurité des données, parcimonie et spécificité. De nouveaux outils d'IA pour lutter contre des problèmes persistants.

William Ritchie*[1]

[1]Institut de Génétique Humaine, Centre National de la Recherche Scientifique (CNRS), Université de Montpellier, Montpellier, France

Résumé

Les données biologiques, surtout génomiques coutent cher à produire et sont difficiles à partager à cause de problèmes d'anonymisation. Dans le domaine médical à ce problème s'ajoute celui des données manquantes à cause de désistements ou de plans d'expérimentations qui sont contraints financièrement. Dans ma présentation je vais montrer comment ces limitations sur les données disponibles peuvent être en partie contournées en utilisant des méthodes d'Intelligence Artificielle et des heuristiques bio-inspirées.

* Intervenant

Communications orales

Dissecting microbe-host relationships via interface-resolved protein interaction networks

Lou Bergogne[1], Jaime Fernandez-Macgregor[1,2], Mégane Boujeant[1], Renaud Vincentelli[2], Christine Brun[1], and Andreas Zanzoni*[1]

[1]Theories and Approaches of Genomic Complexity – Institut National de la Santé et de la Recherche Médicale - INSERM, Aix-Marseille Université - AMU – France

[2]Architecture et fonction des macromolécules biologiques – Aix Marseille Université, Centre National de la Recherche Scientifique – France

Résumé

Pathogens like bacteria and viruses can perturb the host protein interaction network to subvert cellular processes to their own benefit. Protein interaction detection screens show that both bacterial and viral proteins preferentially target protein that are central in the host-cell network. This suggests that targeting proteins with important topological roles is key to perturb the host cell functional organization. Despite the insights generated by these studies, we still have a partial knowledge on how perturbations mediated by pathogen proteins propagate through the host network to trigger the observed subversion. How could pathogens interfere with the host protein interactions at the molecular level?
To address this question, we have designed a strategy that integrates protein-protein interaction data with interface prediction and network analysis. In this framework, we have developed *mimic*INT, a computational method that can infer *bone fide* interfaces using known domain-domain and motif-domain interaction templates, yielding an *in vitro* validation rate up to 70%. More recently, we have focused our attention on bacteria-host protein interactions, and implemented the Mentha workflow, which integrates interface predictions with experimentally identified interactions to generate interface-resolved networks. By doing so, we were able to suggest a putative interface for a significant number of interactions between bacterial proteins and those of three hosts (human, mouse and rat). This dataset will be soon available to the community through the *Bact*Mentha database.

*Intervenant

Benchmarking Data Leakage on Link Prediction in Biomedical Knowledge Graph Embeddings

Galadriel Briere*[1], Thomas Stosskopf[1,2], Benjamin Loire[1,3], and Anaïs Baudot[1]

[1]Aix Marseille Univ, INSERM, MMG, Marseille, France – Aix Marseille Univ, INSERM, MMG, Marseille, France – France

[2]Theories and Approaches of Genomic Complexity – Aix Marseille Université, Institut National de la Santé et de la Recherche Médicale – France

[3]Servier IRIS, Saclay, France – Servier, Paris-Saclay – France

Résumé

In recent years, Knowledge Graphs (KGs) have gained significant attention for their ability to organize complex biomedical knowledge into entities and relationships. Knowledge Graph Embedding (KGE) models facilitate efficient exploration of KGs by learning compact data representations. These models are increasingly applied to biomedical KGs for link prediction, for instance to uncover new therapeutic uses for existing drugs.

While numerous KGE models have been developed and benchmarked for link prediction, existing evaluations often overlook the critical issue of data leakage. Data leakage leads the model to learn patterns it would not encounter when deployed in real-world settings, artificially inflating performance metrics and compromising the overall validity of benchmark results. In machine learning, data leakage can arise when (1) there is inadequate separation between training and test sets, (2) the model leverages illegitimate features, or (3) the test set does not accurately reflect real-world inference scenarios.

In this study, we implement a systematic procedure to control train-test separation for KGE-based link prediction and demonstrate its impact on models' performance. In addition, through permutation experiments, we investigate the potential use of node degree as an illegitimate predictive feature, finding no evidence of such leveraging. Finally, by evaluating KGE models on a curated dataset of rare disease drug indications, we demonstrate that performance metrics achieved on real-world drug repurposing tasks are substantially worse than those obtained on drug-disease indications sampled from the KG.

*Intervenant

Detect de novo expressed ORFs in transcriptomes with DESwoMAN

Anna Grandchamp*[1]

[1]Aix Marseille University, INSERM, TAGC,
UMR$_S$1090, $Marseille, France--CentredeRechercheInserm--France$

Résumé

De novo gene emergence describes the process by which new genes arise from mutations in previously non-coding genomic regions. Before becoming fixed in a species, newly emerged open reading frames (neORFs) undergo significant turnover within their species of origin. Studying these early stages of de novo gene emergence is essential for understanding the mechanisms that enable gene formation from scratch. Although recent software can detect and validate the {de novo} emergence of genes that are fixed in genomes, no software currently exists to extract newly expressed ORFs that are invisible to standard annotation methods or to analyze their mutations and fixation patterns within and across species.

We present DESwoMAN : De novo Emergence Study With Outgroup MutAtioNs, a software tool designed to: (1) detect newly expressed ORFs (neORFS) in transcriptomes, (2) filter neORFs with no homology to outgroup genes, and (3) search for homologous sequences syntenic to neORFs in outgroup genomes (+ optionaly transcriptomes) and extract mutations in coding features between homologs. We applied **{**DESwoMAN} to two different setups, using human and fruitfly as query species, and tested two distinct strategies. Here, we present the main outputs and compare the differences between the options.

*Intervenant

Exonic enhancers are a widespread class of dual-function regulatory elements

Jean-Christophe Mouren*[1], Magali Torres , Antoinette Van Ouwerkerk , Iris Manosalva , Frederic Gallardo , Salvatore Spicuglia , and Benoit Ballester[2]

[1]INSERM, UMR1090 TAGC, Marseille, France – Institut National de la Santé et de la Recherche Médicale – France

[2]INSERM, UMR1090 TAGC, Marseille, France – Institut National de la Santé et de la Recherche Médicale – France

Résumé

Exonic enhancers (EEs) occupy an under-appreciated niche in gene regulation. By integrating transcription factor binding, chromatin accessibility, and high-throughput enhancer-reporter assays, we demonstrate that many protein-coding exons possess enhancer activity across species. These EEs exhibit characteristic epigenomic signatures, form long-range interactions with gene promoters, and can be altered by both nonsynonymous and synonymous variants. CRISPR–mediated inactivation demonstrated the involvement of EEs in the cis-regulation of host and distal gene expression. Through large-scale cancer genome analyses, we reveal that EE mutations correlate with dysregulated target-gene expression and clinical outcomes, highlighting their potential relevance in disease. Evolutionary comparisons show that EEs exhibit both strong sequence constraint and lineage-specific plasticity, suggesting that they serve ancient regulatory functions while also contributing to species divergence. Our findings redefine the landscape of functional elements by establishing EEs as a component of gene regulation, while revealing how coding regions can simultaneously fulfill both protein-coding and cis-regulatory roles.

bioRxiv : https://doi.org/10.1101/2025.03.27.645641

*Intervenant

Tracking down microbial primary producers from the serpentinite-hosted Prony Bay hydrothermal field via metagenomics

Rabja Popall*[1], Melanie Hennart[1], Sophie Marre[2], Pierre Peyret[2], Roy Price[3], Marianne Quéméneur[1], Anne Postec[1], and Gaël Erauso[1]

[1]Institut méditerranéen d'océanologie (MIO) – CNRS : UMR7294, Université du Sud Toulon - Var, Institut de recherche pour le développement [IRD] : UMR235, Aix Marseille Université – France
[2]Microbiologie Environnement Digestif Santé – Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement, Université Clermont Auvergne – France
[3]Stony Brook University – États-Unis

Résumé

In alkaline hydrothermal systems, serpentinization produces geochemical conditions that are very challenging for life. These conditions are believed to reflect the environment in which life might have emerged on Early Earth and potentially on other planetary bodies (1). Serpentinite-hosted ecosystems are thus important analogs for the origin of life (2). However, their development and functioning remain poorly understood. This especially concerns the base of the trophic network: the hyperalkaline milieu produced by serpentinization leads to the precipitation of CO2, immobilizing the only known carbon source used in primary production (3).
Instead of CO2, serpentinite-hosted primary producers might rely on acetate, formate and glycine produced in abiotic reactions linked to serpentinization, or on bicarbonate redissolved from the hydrothermal chimneys (4). In the present study, we assessed whether these compounds could be used by microbial communities from the shallow Prony Bay hydrothermal field. Screening five environmental metagenomes obtained from hydrothermal samples collected along a transect from land to ocean, we identified - based on genetic capabilities - five potential primary producers with a metabolic profile tailored to bicarbonate, formate and acetate uptake in the serpentinite-hosted surface environment. We hypothesize that the base of Prony Bay's trophic network is driven by metabolic interactions between interdependent autotrophs and heterotrophs.

References

1. Schwander L, Brabender M, Mrnjavac N, Wimmer JLE, Preiner M, Martin WF. Serpentinization as the source of energy, electrons, organics, catalysts, nutrients and pH gradients for the origin of LUCA and life. Front Microbiol. 2023;14.

2. Sojo V, Herschy B, Whicher A, Camprubí E, Lane N. The Origin of Life in Alkaline Hydrothermal Vents. Astrobiology. 2016;16:181–97.

*Intervenant

3. Schrenk MO, Brazelton WJ, Lang SQ. Serpentinization, Carbon, and Deep Life. Rev Mineral Geochem. 2013;75:575–606.
4. Popall RM, Postec A, Lecoeuvre A, Quéméneur M, Erauso G. Metabolic challenges and key players in serpentinite-hosted microbial ecosystems. Front Microbiol. 2023;14:1197823.

Communications posters

Expertise et services proposés par la plateforme CiBi du CRCM

Ghislain Bidaut*[1], Samuel Granjeaud[1], Benoît Goutorbe[1], Julien Vernerey[1], Eugénie Lohmann[1], Paraskevi Kousteridou[1], and Pierre Bertrand[1]

[1]Centre de Recherche en Cancérologie de Marseille – Aix Marseille Université, Institut Paoli-Calmettes, Institut National de la Santé et de la Recherche Médicale, Centre National de la Recherche Scientifique, Centre National de la Recherche Scientifique : UMR7258, Institut National de la Santé et de la Recherche Médicale : U1068, Institut Paoli-Calmettes : UMR7258, Aix Marseille Université : UM105 – France

Résumé

La plateforme bioinformatique CiBi du CRCM a pour mission principale de coordonner et de réaliser des analyses bioinformatiques, notamment le traitement de données issues des technologies de séquençage de nouvelle génération (NGS), telles que le **ChIP-Seq, RNA-Seq, DNA-Seq,** ainsi que l'**intégration multi-omique**.
Elle intervient également dans l'analyse de données issues du **Single Cell RNA-seq**. La plateforme contribue à l'exploitation des données de cytométrie et de protéomique, et accompagne l'émergence et l'intégration de nouvelles technologies, notamment la **transcriptomique spatiale** via les plateformes 10X Visium et Visium HD.

Elle participe également activement à la création et à la gestion de **bases de données** biomédicales, appliquées au soin et utilisées à l'Institut Paoli-Calmettes (sous framework Django).
Un des objectifs clés est le développement de programmes d'analyse bioinformatique, tels que **CellRegulomiX**, dédié à l'intégration de données multi-omiques et à l'étude du rôle des facteurs de transcription dans des expérimentations à haut débit en biologie.

*Intervenant

Gammaproteobacteria Operons for polysaccharide utilization: from Algae to a Larger Scope (GOALS project)

Matthieu Boulinguiez*[1] and Nicolas Terrapon[1]

[1]Architecture et fonction des macromolécules biologiques – Aix Marseille Université, Centre National de la Recherche Scientifique, Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement – France

Résumé

The degradation of polysaccharides from marine microbial biofilms and the extracellular matrices of macroalgae suggest a wealth of valuable compounds for health and industry (1). Recent progress in understanding polysaccharide degradation took advantage of bacterial operons termed Polysaccharide Utilization Loci (PUL) found in the Bacteroidota phylum. Consequently the Carbohydrate-Active enZyme (CAZy) database (2) developed the PULDB to present to the community loci from the literature and predictions for annotated genomes. The Gammaproteobacteria phylum comprises diverse organisms from pathogens to commensals, in terrestrial or aquatic microbiomes. Among these, some operonic CAZyme arsenals have been identified within the order Alteromonadales (3), and to a lesser extent in Oceanospirillales, Vibrionales and Cellvibrionales. The GOALS project aims at (i) massive genome annotation in these Gammaproteobacteria orders, (i) exploit this data for the large-scale prediction of their PULs and (iii) identify conserved protein of unknown function as targets for biochemical assays to discover novel enzymatic activities and families.
The first step of the GOALS project was to select diverse and relevant genomes to analyze. For this purpose, we developed a tool to explore the flood of genomes (of variable quality) available at NCBI, and prioritize candidates increasing our taxonomic diversity, with a high degree of assembly and a marine CAZyme-rich/diverse profile. Nearly 150 Alteromonadales genomes have been collected, we annotated their CAZymes with HMMer & BLAST approaches, through a semi-manual expert curation. The second step is the development of a prediction algorithm that will be evaluated on experimental data. For this purpose, we survey the literature to identify 15 PULs, used as references. We are additionally analyzing transcriptomic data generated by the "Station Biologique de Roscoff" to produce more experimental/reference PULs to train our algorithm. We will then apply this algorithm to predict PULs in the 445 Gammaproteobacteria genomes with annotated CAZymes. From these predictions, the final step consists in identifying proteins of unknown function, conserved in these PULs and beyond. This approach uses a sequence similarity network to cluster them into groups. Functional domains are then extracted by combining pairwise alignments, and statistical/structural analyses are performed to discriminate promising candidates (novel CAZymes) from other PUL actors (sulfatases, transporters, regulators), or unrelated genes.

*Intervenant

Knowledge graph to dissect genotype phenotype associations

Florence Ghestem*[1,2], Anne-Louise Leutenegger[1,2], and Anaïs Baudot[3,4]

[1]Centre de recherche en épidémiologie et santé des populations – Université de Versailles Saint-Quentin-en-Yvelines, Assistance publique - Hôpitaux de Paris (AP-HP), Hôpital Paul Brousse, Institut National de la Santé et de la Recherche Médicale, Université Paris-Saclay – France

[2]NeuroDiderot Inserm U1141 – Université Paris Cité – France

[3]Aix Marseille Univ, INSERM, MMG, 13385, Marseille, France – Institut National de la Santé et de la Recherche Médicale - INSERM, Aix-Marseille Université - AMU, Aix Marseille Univ, INSERM, MMG, Marseille Medical Genetics, Marseille, France – France

[4]Barcelona Supercomputing Center (BSC), 08034 Barcelona, Spain . – Espagne

Résumé

Defining the mechanisms underlying complex traits and phenotypes requires a comprehensive understanding of their genetic basis, as well as how genes interact with environmental and lifestyle factors. Population-based prospective cohorts provide a valuable resource for such research. Such cohorts often collect extensive phenotypic data (*e.g.* laboratory tests, health outcomes thanks to linked hospital and primary care records), as well as omics data. However, they typically include participants with a wide range of phenotypes, making it difficult to define subgroups of individuals that are phenotypically homogeneous and/or clinically meaningful, which overall limits traditional genome-wide association studies (GWAS). We aim here to develop a novel graph-based methodology to identify genotype-phenotype associations. Our method first represents data as a graph, in which nodes correspond to the variables present in the cohort (*e.g.*, participants, phenotypes) and edges represent the relationship between the nodes. These edges can connect either the same type of nodes (*e.g.*, participant-participant interactions) or different types of nodes (bipartite interactions connecting *e.g.* patient to SNPs).

We applied our methodology to data from the GOLD project, which includes comprehensive information on medical conditions, drug consumption, demographic characteristics, and genotypes of 10,000 participants. The graph contains 4 types of nodes (participant, drug, SNP, phenotype) and 10 types of edges. Edges can have attributes, for example, the number of reimbursements for a drug, the genotype of a specific SNP for a patient, or participant-participant similarity.
This knowledge graph representation allows us to use tools from graph theory, including clustering algorithms, random walks, and deep learning-based graph representation methods. We expect to detect weak genotype-phenotype association signals that could not be detected by GWAS.

*Intervenant

Met'Connect: Bioinformatics necessary tool to explore Tumor Metabolism

Paraskevi Kousteridou*[1] and Pierre Bertrand*[2]

[1]Centre de Recherche en Cancérologie de Marseille – Centre de Recherche en Cancérologie de Marseille (CRCM), INSERM U1068, Institut Paoli-Calmettes, Aix-Marseille Université, CNRS UMR7258, Marseille, France. – France

[2]Centre de Recherche en Cancèrologie de Marseille – Centre de Recherche en Cancérologie de Marseille (CRCM), INSERM U1068, Institut Paoli-Calmettes, Aix-Marseille Université, CNRS UMR7258, Marseille, France. – France

Résumé

Tumor metabolism is an a key player in tumor progression and aggressiveness field in cancer biology, in both clinical and fundamental research. Serving as a dynamic diagnostic base, tumor metabolism provides invaluable insights into the intricate interplay of metabolites and their profound implications for patho-physiological and patho-biochemical processes, particularly in cancer progression. As a result, there is an emerging need and growing interest for tailored bioinformatics approaches for metabolism analysis.

Met'Connect is a unique Cancerolopole-founded structuring action launched to address needs in metabolism cancer research, in both wet lab and dry lab, sharing the expertise of a multidisciplinary network of researchers in South of France. A distinctive feature of Met'Connect lies in its seamless integration with multi-omics datasets, providing researchers with the flexibility to combine various types of omics data (e.g. transcriptomics, metabolomics, lipidomics, proteomics) for a comprehensive
and holistic understanding of biological processes. This integration enhances the platform's ability to unravel the complex interplay between different molecular layers and uncover novel insights into tumor metabolism, ultimately advancing our understanding of cancer biology and guiding the development of novel targeted therapies combining metabolic targeting with current chemotherapies.
The aim of Met'Connect is to analyze different types of omics data to discover metabolic profiles of tumor samples based on specific conditions (e.g. treatment, tumor stages etc.). The structuring action delivers results in user-friendly HTML reports and Excel formats, making them accessible and usable for tailored research needs.

Addressing the growing demand for accessible multi-omics analysis in PACA region, Met'Connect arises as a specialized structure uniquely poised to address the need for accessible multi-omics analysis, specifically tailored for researchers seeking to decipher the metabolic intricacies of tumors. Our services are available upon request for any researcher particularly with no expertise in tumor metabolism and strengthen collaborations in the southern regions, with the aim to expand collaborations at inter-regions and national scales.

*Intervenant

Bioluminescence regulation in Photobacterium phosphoreum ANT2200: the role of the tfoX gene

Capucine Le Cam Ligier*[1], Gwenola Simon[1], Corinne Valette[1], Elisa Lefeuvre[2], Florian Haitz[1], Marc Garel[1], Léa Girolami[3], Charlotte Berthelier[4], and Laurie Casalot[1]

[1]Institut méditerranéen d'océanologie – Institut de Recherche pour le Développement, Aix Marseille Université, Institut National des Sciences de l'Univers, Université de Toulon, Centre National de la Recherche Scientifique, Institut de Recherche pour le Développement : UMR$_D$235, $AixMarseilleUniversité : UM110, InstitutNationaldesSciencesdel'Univers : UMR7294, CentreNationaldelaRechercheScientifique : UMR7294, UniversitédeToulon : UMR7294 − − France$

[2]Université de Bordeaux – Université de Bordeaux (Bordeaux, France) – France

[3]Lycée Marie Curie, Marseille – Lycée Marie Curie – France

[4]Station biologique de Roscoff = Roscoff Marine Station – Sorbonne Universite, Centre National de la Recherche Scientifique – France

Résumé

The bioluminescence reaction in bacteria is an essential ecological process in deep-sea environments (1,2). The bacterial bioluminescence is predominantly controlled by the lux operon, a genetic system encoding the genes essential for light production (3,4). While the lux operon itself is well characterized, various genetic regulatory factors throughout the genome that influence overall light emission remain poorly understood (5). In this study, we investigated novel regulatory elements affecting bioluminescence by comparing the genome and transcriptome of two strains of Photobacterium phosphoreum ANT2200: a wild-type strain Lum and a spontaneous DimLum mutant with significantly reduced luminescence. Comparative analyses identified the gene tfoX as a potential key factor in modulating light emission. tfoX encodes a transcription factor known for its role in regulating natural competence in Gram-negative bacteria (6). However, its potential involvement in bioluminescence regulation had not been previously reported. Using bacterial conjugation, we introduced the wild-type tfoX gene into the DimLum strain, successfully restoring luminescence to wild-type levels. Furthermore, ongoing high-pressure incubation experiments aim to assess how deep-sea environmental conditions influence the growth, luminescence intensity, and emission wavelength (7, 8, 9) in the three strains. By characterizing Lum, DimLum, and the tfoX conjugant under identical conditions, we hope to understand how tfoX plays a crucial role in regulating light emission in P. phosphoreum. This study uncovers previously unknown aspects of bioluminescence control and provides new insights into light and growth conditions.

References

(1) S. H. D. Haddock, M. A. Moline, and J. F. Case, 'Bioluminescence in the Sea', *Annu. Rev. Mar. Sci.*, vol. 2, no. 1, Art. no. 1, 2010, doi: 10.1146/annurev-marine-120308-081028.

*Intervenant

(2) L. Tanet, S. Martini, L. Casalot, and C. Tamburini, 'Reviews and syntheses: Bacterial bioluminescence – ecology and impact in the biological carbon pump', *Biogeosciences*, vol. 17, no. 14, Art. no., 2020, doi: 10.5194/bg-17-3757-2020.

(3) E. A. Meighen, 'Bacterial bioluminescence: organization, regulation, and application of the *lux* genes', *FASEB J. Off. Publ. Fed. Am. Soc. Exp. Biol.*, vol. 7, no. 11, pp. 1016–1022, 1993, doi: 10.1096/fasebj.7.11.8370470.

(4) J. Mancini, M. Boylan, R. Soly, S. Ferri, R. Szittner, and E. Meighen, 'Organization of the *lux* genes of *photobacterium phosphoreum*', *J. Biolumin. Chemilumin.*, vol. 3, no. 4, Art. no. 4, 1989, doi: 10.1002/bio.1170030407.

(5) E. A. O'Grady and C. F. Wimpee, 'Mutations in the lux operon of natural dark mutants in the genus Vibrio', *Appl. Environ. Microbiol.*, vol. 74, no. 1, pp. 61–66, 2008, doi: 10.1128/AEM.01199-07.

(6) L. C. Metzger, N. Matthey, C. Stoudmann, E. J. Collas, and M. Blokesch, 'Ecological implications of gene regulation by TfoX and TfoY among diverse *Vibrio* species', *Environ. Microbiol.*, vol. 21, no. 7, pp. 2231–2247, 2019, doi: 10.1111/1462-2920.14562.

(7) B. Al Ali *et al.*, 'Luminous bacteria in the deep-sea waters near the ANTARES underwater neutrino telescope (Mediterraean Sea)', *Chem. Ecol.*, vol. 26, no. 1, Art. no. 1, 2010, doi: 10.1080/02757540903513766.

(8) S. Martini *et al.*, 'Effects of Hydrostatic Pressure on Growth and Luminescence of a Moderately-Piezophilic Luminous Bacteria *Photobacterium phosphoreum* ANT-2200', *PLOS ONE*, vol. 8, no. 6, Art. no. 2013, doi: 10.1371/journal.pone.0066580.

(9) M. Garel *et al.*, 'Pressure-Retaining Sampler and High-Pressure Systems to Study Deep-Sea Microbes Under in situ Conditions', *Front. Microbiol.*, vol. 10, p. 453, 2019, doi: 10.3389/fmicb.2019.00453.

KGATE, an Autoencoder Training Environment for exploring and benchmarking Knowledge Graph Embedding models

Benjamin Loire*[1,2], Galadriel Brière[2], and Anaïs Baudot[2,3]

[1]Laboratoire Servier – Institut de Recherches Internationales Servier [Suresnes] – France
[2]Marseille medical genetics - Centre de génétique médicale de Marseille – Aix Marseille Université, Institut National de la Santé et de la Recherche Médicale – France
[3]CNRS – CNRS Marseille – France

Résumé

Background

Knowledge Graphs (KGs) are data structures allowing the organisation of massive heterogeneous data into a common framework. Different machine learning algorithms have been developed to learn from these data structures and extract patterns leading to data-driven discoveries. Knowledge Graph Embedding (KGE) is a machine learning technique that projects the entities and relationships of a KG into continuous vector spaces. One of the main applications of KGE, especially in biomedical contexts, is the completion of KG by predicting novel relationships. KGE models can be formalised as autoencoders, where an encoder embeds the graph into a latent space and a decoder reconstructs the original graph from this latent representation.
The two main libraries for KGE in Python are Pytorch Geometric (PyG) and TorchKGE. PyG is a versatile deep learning graph embedding library with powerful encoders, but is targeted towards advanced users. TorchKGE, on the other hand, offers a simple and comprehensive API to train KGE decoders. However, TorchKGE does not implement encoders. To combine PyG's deep learning capabilities with TorchKGE's ease of use, we propose KGATE, a Knowledge Graph Autoencoder Training Environment.

KGATE is a comprehensive KG autoencoder framework that integrates PyG's powerful encoders with TorchKGE's specialised decoder capabilities, creating an easy-to-use and unified environment for KGE exploration and benchmarking. KGATE can train and test a model in as few as four lines of code, while providing an API for in depth analyses. In addition, KGATE implements functions to detect and control data leakage during the training process, hyperparameter optimisation, and training of a triplet classifier to validate the relevance of predicted relationships.

KGATE aims to provide data scientists with a unified framework to explore KGE models. The ease of use reduces the requirement for advanced machine learning knowledge and the need for multiple complementary libraries.

*Intervenant

Transcriptomic profile characterization at the single-cell level to uncover regulatory mechanisms responsible for malaria outcomes

Loréna Quatreville*[1], Mathieu Adjemout[1], Charlyne Gard[2], Magali Torres[1], Camille Cohen[3], Brigitte Tunamo , Lionel Spinelli[4], Sandrine Eveline Nsango , Antoine Claessens[5], and Sandrine Marquet[1,6]

[1]Aix-Marseille Université, INSERM, TAGC, UMR 1090, MarMaRa Institute, Marseille, France. – Aix-Marseille Université - AMU, Institut National de la Santé et de la Recherche Médicale - INSERM – France

[2]Transcriptomics and Genomics of Marseille-Luminy – Centre de Recherche Inserm, Aix-Marseille Université - AMU – France

[3]INMG-PGNM – Institut NeuroMyoGene (INMG-PGNM) – France

[4]CIML – Aix Marseille University, CNRS, INSERM, CIML, Marseille, France – France

[5]LPHI - Laboratory of Pathogen and Host Immunity (LPHI) – Centre National de la Recherche Scientifique, Université de Montpellier – France

[6]Aix Marseille Université, CNRS, Marseille 13009, France – Aix-Marseille Université - AMU, CNRS – France

Résumé

In 2023, malaria was still responsible for more than 250 million cases worldwide and accounted for almost 600 000 deaths the same year (WHO). In this context, this disease is still a major public health concern, especially in African countries accounting for around 95 % of malaria cases and deaths.
Malaria is given by a protozoan parasite, *Plasmodium falciparum* for the most common, transferred from an infected individual to a heathy one by an *Anophele* mosquitoe's bite. During the erythrocytic asexual cycle of the parasite, around 25 % infected individuals develop mild symptoms such as fever, less than 1 % having severe malaria possibly leading to death. The 75 % remaining are asymptomatic individuals, developing chronic infection and therefore constituting the parasite reservoir. The wide variety of host responses can be explained by environmental parameters, the host and the parasite genomic backgrounds.

In our study, we aim to use a multi-omic approach to uncover regulatory mechanisms that are responsible for the different malaria outcomes.

The study we are conducting is based on a longitudinal cohort of Cameroonian children either developing mild symptoms or remaining asymptomatic.
Using a single-cell RNA-Seq experiment, we want to characterize the transcriptomic profile of PBMCs from children being symptomatic or having a chronic infection at two different time-points. The first time-point is at the beginning of the infection when children become

*Intervenant

positive to *P. falciparum*. The second sampling occurs either at the development of symptoms, or at the end of the study (2 months later) if the individuals remained asymptomatic. 16 samples (11 individuals) have been sequenced. Due to cell viability variability, we have the two time-points information for 5 individuals out of 11. The libraries and sample quality are high with a recovery of 10 000 cells per individual in average and around 35 000 mean read per cell. The sequencing saturation is around 80% for 12 individuals from the same batch, a bit less (60%) for the first 4 individuals sequenced. We employed an existing single-cell RNA-Seq analysis pipeline using the Seurat package, which we adapted and customized to suit the specific requirements of our dataset and research objectives. When comparing the two time-points, the main difference seems to lie in the monocyte cell-type in terms of cell number as well as in transcriptomic profile while further analyses are needed to corroborate these observations. When comparing monocytes from the second time-point of chronic individuals with symptomatic ones, it appears that symptoms severity and variability is associated to very distinct transcriptomic profiles while chronic individual monocytes seem more homogeneous. This suggests that protection against malaria would be due to a more stereotyped response while sensitive individuals would have very unique responses possibly reflecting the different symptomatic profiles.

Predictions of human micropeptide functions from an interactome study

Mathilde Slivak*[1], Sebastien A. Choteau[1], Lionel Spinelli[1], Andreas Zanzoni[1], and Christine Brun[2]

[1]Theories and Approaches of Genomic Complexity – Aix Marseille Université, Institut National de la Santé et de la Recherche Médicale – France
[2]CNRS – CNRS Marseille – France

Résumé

Short Open Reading Frames (sORFs) are ubiquitous genomic elements that have been overlooked for years, essentially due to their short length (< 100 residues) and the use of alternative start codons other than AUG. However, some may encode functional micropeptides whose number (_˜7000 to several hundred thousand in human) and functions remain uncertain.
Here we propose a system approach to determine the functions of human micropeptides in monocytes and skeletal muscle cells. Based on metamORF, the sORF database that we previously developed (1), we selected 4522 and 590 micropeptides expressed in monocytes and muscle cells, respectively.

First, using mimicINT, a method for large-scale protein-protein interaction interface prediction that we developed recently (2), we predicted the interactions of 1810 micropeptides with 11416 canonical proteins in monocytes and 252 micropeptides with 1588 canonical proteins in muscle cells. A thorough analysis of the interaction interfaces detected in micropeptides underlines their trend to be phosphorylated.

Second, by joining these micropeptide-canonical protein interactions with the human interactome, we built the first cell-specific micropeptide containing-interactome networks to date, constituted of 113 722 interactions between 12414 proteins, among which 1810 microproteins in monocytes, and 172 015 interactions between 16 681 proteins, among which 252 microproteins, in muscle cells. Decomposing these interactomes in protein clusters using our OCG algorithm (3), we then predicted the function of the micropeptides based on their belonging to protein clusters of known cellular functions.

Our results suggest that the majority of micropeptides are involved in key biological functions, including immune response, regulatory functions, metabolism, and signaling. Overall, the diversity in the predicted functions of the micropeptide underlines the prevalence of their role in different biological mechanisms, suggesting that they are major regulatory actors (4). We are currently generating a brain cell micropeptide-containing interactome network for further comparison analyses in order to investigate in detail the commonalities and cell-specificities of the predicted functions of the micropeptides.

*Intervenant

(1) Choteau S., et al. (2021) Database, baab032.

(2) Choteau S., et al. (2022) bioRxiv, doi: 10.1101/2022.11.04.515250

(3) Becker E., et al. (2012) Bioinformatics. 28:84-90.

(4) Slivak M., et al. (2024) bioRxiv, 10.1101/2024.06.10.598216

Sponsors


Réseau
MERIT
MetiER en
bioinformaTique


sfbi
société française de bioinformatique


mio
Institut Méditerranéen
d'Océanologie