Le pôle analyse: analyse de données pour l'appui à la recherche

Jawad Abdelkrim*[1]

[1] MNHN DGD-REVE, UMS 2700 Acquisition et Analyse de Données pour l'histoire naturelle (2AD)

Le pôle analyse de l'UMS2700, Acquisition et Analyse de Données pour l'Histoire naturelle (2AD), a pour objectif d'offrir un appui à la recherche pour le Muséum et ses partenaires en réunissant des compétences dans le domaine de l'analyse de données. Bien que les chercheurs du Muséum maîtrisent des systèmes divers, l'exploitation et la valorisation des données restent parfois des facteurs limitants.

C'est pourquoi le pôle analyse propose d'accompagner la communauté du Muséum par la participation ponctuelle ou soutenue à l'analyse de données, le développement méthodologique, la formation, l'aide à la conception de projets et la collaboration. Les domaines de compétence du pôle sont les biostatistiques, l'analyse en imagerie, la génétique/génomique fonctionnelle et évolutive, ainsi que le calcul à haute performance.

Nous présenterons les membres et compétences du pôle, les outils et les services communs proposés ainsi que les ateliers/formations mis en place, en insistant sur les aspects bioinformatique. Cette rencontre sera l'occasion de situer nos ressources et d'identifier nos liens actuels et potentiels avec la communauté bioinformatique et les utilisateurs du Muséum.

---

* Intervenant

Quels processus expliquent la diversité au sein des espèces

Guillaume Achaz*[1]

[1] UMR 7206 Département Homme et environnement, MNHN

Comment peut-on expliquer la diversité au sein des espèces en 2019 ? Dès l'avènement de la théorie neutraliste (1968), R Lewontin note que malgré les prédictions claires de cette théorie, la diversité moléculaire observée n'est pas corrélée linéairement avec la taille des populations (1972). Malgré tout, les méthodes d'inférence démographique reposant sur cette théorie font des prédictions empiriques très censées, au regard des nos connaissances biologiques... pourvu que la taille de population soit remplacée par une taille dite "efficace". Bref nous utilisons un modèle clairement erroné qui pourtant produit des prédictions "justes" (acceptables). Alors quelles conclusions tirer de tout cela ? Qu'est-ce que cette taille efficace ? Quel est son sens et que nous dit-elle sur les processus qui sculptent la diversité ? Existe-t-il un modèle plus juste qui expliquerait la diversité observée et permettrai de faire des inférences démographiques cohérentes ? Je discuterai de ces questions évolutives fondamentales en combinant des approches de biologie, de modélisation et de bioinformatique.

Caractériser le génome et le protéome de grégarines pour comprendre la diversification des Apicomplexes et leur adaptation à la vie parasitaire

Julie Boisard * [1], Evelyne Duvernois-Berthet [2], Loïc Ponger [3], Isabelle Florent [1]

[1] MCAM, Molécules de Communication et Adaptations des Microorganismes – Muséum National d'Histoire Naturelle (MNHN), CNRS : UMR7245 – France
[2] Evolution des régulations endocriniennes (ERE) – CNRS : UMR7221 – 7 Rue Cuvier 75231 PARIS CEDEX 05, France
[3] Structure et Instabilité des Génomes (STRING) – Muséum National d'Histoire Naturelle (MNHN), Inserm : U1154, CNRS : UMR7196 – France

Les Apicomplexes sont un groupe monophylétique d'eucaryotes unicellulaires comprenant 6000 espèces décrites, qui ont toutes adopté un mode de vie parasitaire strict dans une diversité d'hôtes Métazoaires. Alors que les parasites intracellulaires de vertébrés tels *Plasmodium*, Toxoplasma, Eimeria, Cryptosporidium, responsables de pathologies graves (Paludisme, Toxoplasmose, Coccidioses, Cryptosporidioses) sont bien connus sur le plan génomique, la majeure partie de la biodiversité des Apicomplexes, représentée par les grégarines, est méconnue (1 génome pour 1700 espèces décrites).
Les grégarines sont principalement extracellulaires et considérées non pathogènes ; elles ne sont pas cultivables. Des données récentes ont révélé que les génomes connus des Apicomplexes ont subi des réductions massives de gènes à partir de leurs ancêtres libres proto-Apicomplexes, *Chromera* et Vitrella, en lien avec leur adaptation au mode de vie parasitaire.
Afin de pouvoir disposer d'une vision complète de l'histoire évolutive des Apicomplexes et comprendre les mécanismes moléculaires qui ont progressivement permis la mise en place du parasitisme pour ce groupe, leur permettant d'évoluer d'un mode de vie extra à intracellulaire au sein d'une très grande diversité d'hôtes et avec une pathogénicité croissante, nous avons entrepris le séquençage *de* novo du génome de trois grégarines ; les premiers résultats de ces travaux seront présentés.

*Intervenant

Basecaller training, and the bottleneck to bring ONT to non conventional models

Nicolas Buisine*[1]

[1] UMR CNRS 7221 Physiologie Moleculaire et Adaptations

The Nanopore technology sequences DNA by measuring changes of electric current when it passes through a pore separating two ionic solutions. Unfortunately, DNA is sequenced 5 to 6 bases at a time and the game purpose is to translate signal in k-mer space to DNA sequence space. This is carried out by combination of deep learning technologies, that critically rely on appropriately trained models. Although available for a few biological species, these models are fairly challenging to develop.

Protein Multiple Alignments: Sequence-based vs Structure-based Programs

Mathilde Carpentier * [1], Jacques Chomilier [2]

[1] Institut de Systématique, Evolution, Biodiversité UMR 7205 – Sorbonne Université UPMC Paris VI, Museum National d'Histoire Naturelle - MNHN (FRANCE), Ecole Pratique des Hautes Etudes, CNRS : UMR7205 – France

[2] Institut de minéralogie, de physique des matériaux et de cosmochimie (IMPMC) – Museum National d'Histoire Naturelle, Institut de recherche pour le développement [IRD] : UR206, Sorbonne Universite : UM120, Centre National de la Recherche Scientifique : UMR7590 – Tour 23 - Barre 22-23 - 4e étage - BC 115 4 place Jussieu 75252 PARIS, France

Motivation: Multiple sequence alignment programs have proved to be very useful and have already been evaluated in the literature yet not alignment programs based on structure or both sequence and structure. In the present article we wish to evaluate the added value provided through considering structures. **Results:** We compared the multiple alignments resulting from 25 programs either based on sequence, structure, or both, to reference alignments deposited in five databases (BALIBASE 2 and 3, HOMSTRAD, OXBENCH and SISYPHUS). On the whole, the structure-based methods compute more reliable alignments than the sequence-based ones, and even than the sequence+structure-based programs whatever the databases. Two programs lead, MAMMOTH and MATRAS, nevertheless the performances of MUSTANG, MATT, 3DCOMB, TCOFFEE+TM_ALIGN and TCOFFEE+SAP are better for some alignments. The advantage of structure-based methods increases at low levels of sequence identity, or for residues in regular secondary structures or buried ones. Concerning gap management, sequence-based programs set less gaps than structure-based programs. Concerning the databases, the alignments of the manually built databases are more challenging for the programs.

*Intervenant

Présentation de la Cellule de Soutien Bioinformatique du département AVIV

Evelyne Duvernois-Berthet * [1,2], Loïc Ponger [3,2]

[1] Physiologie Moléculaire et Adaptation (PhyMA) – Muséum National d'Histoire Naturelle (MNHN), CNRS : UMR7221 – France
[2] Cellule de Soutien Bioinformatique du Département AVIV – Muséum National d'Histoire Naturelle (MNHN) – France
[3] Structure et Instabilité des Génomes (STRING) – Muséum National d'Histoire Naturelle (MNHN), Inserm : U1154, CNRS : UMR7196 – France

La Cellule de Soutien Bioinformatique (CSB) a été créée en 2011 afin de répondre à toutes les demandes (bio)informatiques de l'ancien département RDDM du Muséum national d'Histoire naturelle. Depuis, la CSB est intégrée au département AVIV. Elle a pour missions principales l'aide et la formation en (bio)informatique pour les membres du département. Notre expertise va de l'analyse de séquences moléculaires au développement d'outils informatiques. A l'occasion des Rencontres Bioinformatiques du MNHN 2019, nous vous présenterons le fonctionnement de la Cellule ainsi quelques exemples d'études que nous avons pu mener en collaboration avec différentes unités du MNHN.

*Intervenant

Applications bioinformatiques pour l'étude des interactions algues pathogènes

Claire Gachon[*] [1]

[1] Molécules de Communication et Adaptation des Micro-organismes (MCAM) – Museum National d'Histoire Naturelle, Centre National de la Recherche Scientifique : UMR7245 – France

Nouvellement recrutée au Muséum pour assurer la coordination scientifique du pôle analyse de données de l'UMS 2AD (Acquisition et Analyse de Données pour l'histoire naturelle), je suis également rattachée au MCAM et à la Scottish Association for Marine Science. Je présenterai dans cet exposé quelques exemples d'approches (bio)informatiques sur les maladies des algues brunes, des algues rouges et des diatomées, à la croisée des chemins entre l'écologie, la génomique comparée et l'aquaculture.

Dans un premier temps, je traiterai de l'utilisation du séquençage en single-cell pour accélérer la description de nouveaux agents pathogènes pertinents pour la dynamique du phytoplancton, et de la façon dont nous en tirons de nouveaux modèles pour la génomique comparative des oomycètes. Je présenterai un exemple d'annotation fonctionnelle du génome de l'algue brune *Ectocarpus siliculosus*, notamment de deux famille multigéniques potentiellement impliquées dans la reconnaissance du non-soi. Enfin, j'introduirai quelques projets de transcriptomique sur des interactions hôtes-pathogènes et de GWAS (Genome-Wide Association Study) sur l'algue brune d'intérêt commercial *Saccharina latissima*.

[*] Intervenant

Bases omiques de l'étude de l'écologie d'Aphanizomenon gracile et de sa cyanosphère

Sébastien Halary * [1]

[1] Molécules de Communication et Adaptation des Micro-organismes (MCAM) – Museum National d'Histoire Naturelle, Centre National de la Recherche Scientifique : UMR7245 – France

La fréquence et l'amplitude des épisodes d'efflorescence de cyanobactéries ont fortement augmenté ces dernières années. Parmi elles, le genre *Aphanizomenon* est considéré comme l'un des plus toxiques et a des conséquences particulièrement délétères sur les écosystèmes aquatiques. Si les facteurs abiotiques favorisant ces efflorescences sont bien caractérisés (*e.g.* excès de phosphore, luminosité, température), les facteurs biotiques restent quant à eux mal connus. Pourtant, les cyanobactéries semblent intégrées à un réseau d'interactions étroites avec d'autres micro-organismes, tel qu'il est très difficile d'obtenir des cultures axéniques pour de nombreux taxa. Quelques exemples d'interactions ont été décrits au sein de cyanosphères marines (protection des cyanobactéries contre le stress oxydatif par des bactéries hétérotrophes, échange de précurseurs de voies de biosynthèse), mais les cyanosphères d'eau douce reste peu étudiées et la diversité phylogénétique des membres de ces communautés suggèrent un large spectre d'interactions possibles. Cet exposé présentera les résultats d'une étude de (méta)génomique et de métabolomique comparative de plusieurs cultures non-axéniques d'*A. gracile* isolées d'un même échantillon environnemental et des cyanosphères associées, visant à mieux comprendre les facteurs biotiques régulant la croissance des cyanobactéries lacustres.

*Intervenant

MNHN-Tools: From sequences to phylogeny

Thomas Haschka*[1]

[1] Structure et Instabilité des Génomes (STRING) – Muséum National d'Histoire Naturelle (MNHN)

We present a novel developed suite of tools *MNHN-Tools* that allows us to cluster and build phylogenetic trees from a dataset containing nucleic sequences. As the number of nucleic sequences grow exponentially and faster than Moore's law, the endeavor to find more efficient and flexible algorithms is both crucial and infinite. MNHN-Tools are based on an iterative, adaptive version of the DBSCAN[1] algorithm and are further optimized to take advantage of Single Instruction Multiple Data (SIMD) instructions, Graphics Processing Unit (GPU) computing, multithreading and multinode high performance computing (HPC) architectures. Adapting the density iteratively of the well known density-based algorithm for discovering clusters in large spatial databases with noise (DBSCAN) clustering algorithm allows us to gain insights into the evolution of sequences and to build phylogenetic dependency graphs ultimately leading to trees.

We will further outline how we apply this method to find new families of α-satellite families in old world monkeys[2] and how we benchmark our tool against available classified datasets of α-satellites in humans[3] as well as against the annotated sequences that form a tree of life (Tol) from the SILVA[4] project.

[1] M. Ester et al. KDD96 proceedings: A Density-Based Algorithm for Discovering Clustersin Large Spatial Databases with Noise

[2] L. Cacheux et al. Genome Biology and Evolution 2018(7) p 1837-1851

[3] V. A. Shepelev et al. Genomics Data 2015(5) p 139-146

[4] P. Yilmaz et al. Nucleic Acids Res. 2014 Jan 1; 42(Database issue): D643–D648.

Complex genetic admixture histories reconstructed with Approximate Bayesian Computations

Romain Laurent * [1], Paul Verdu [1]

[1] Éco-Anthropologie – Museum National d'Histoire Naturelle, Université Paris Diderot - Paris 7, Centre National de la Recherche Scientifique : UMR7206 – France

Admixture is a fundamental evolutionary process that has influenced genetic patterns in numerous species. Maximum-likelihood approaches based on allele frequencies and linkage-disequilibrium have been extensively used to infer admixture processes from dense genome-wide datasets, mostly in human populations. Nevertheless, complex admixture histories, beyond one or two pulses of admixture, remain methodologically challenging to reconstruct, especially when large datasets are unavailable. We develop an Approximate Bayesian Computations (ABC) framework to reconstruct complex admixture histories from independent genetic markers. We built the software package *MetHis* to simulate independent SNPs in a two-way admixed population for scenarios with multiple admixture pulses, or monotonically decreasing or increasing admixture at each generation; drawing model-parameter values from prior distributions set by the user. For each simulated dataset, we calculate 24 summary statistics describing genetic diversity and moments of individual admixture fractions. We coupled *MetHis* with existing machine-learning ABC algorithms and investigate the admixture history of hybrid populations. We focus on an African American and a Barbadian population as case studies of human admixture. Results show that Random-Forest ABC scenario-choice, followed by Neural-Network ABC posterior parameter estimation, can distinguish most complex admixture scenarios and provide accurate model-parameter estimations.

*Intervenant

Galaxy-E une plateforme européenne web d'analyse de données dédiée à l'écologie.

Yvan Le Bras*[1]

[1] MNHN DGD-REVE, UMS PatriNat Station de Biologie Marine de Concarneau (MNHN) – Muséum National d'Histoire Naturelle (MNHN) – France

En 2017, dans le cadre du pia 65 millions d'observateurs naissait l'initiative "Galaxy for Ecology" Galaxy-E de son petit nom. Aujourd'hui développée dans le cadre du pôle national de données de biodiversité, Galaxy-E joue un rôle majeur dans

1/ le projet pilote d'operationnalisation des variables essentielles de biodiversité (EBVs) pour la production automatisée de métriques et indicateurs de biodiversité à l'échelle des populations et des communautés dans le cadre de ma participation française à GEO BON,

2/ la mise en place d'un cas d'étude "biodiversité" de l'utilisation du cloud européen open science EOSC, sur les aspects modélisation de niches via le projet EOSC-PILLAR

3/ la mise en place de formations innovantes en e-learning orientées vers les jeunes populations et leur éveil au traitement analytique de données scientifiques dans le cadre du projet Européen ERASMUS+ Gallantries

4/ l'initiative innovante Galaxy-Bricks portée par Vigie-Nature école et permettant à des collégiens d'appréhender dans un environnement dédié les notions de biodiversité, de mathématiques appliquées et d'informatique.

L'équipe PNDB propose de faire un focus sur ce fabuleux environnement qu'est Galaxy en illustrant via ces initiatives les fonctionnalités qui font sont succès et qui pourrait vous permettre de répondre à vos besoins.

---

* Intervenant

GOTIT (Gene Occurrence and Taxa in Integrative Taxonomy)

Florian Malard * [1], Philippe Grison *

[2]

[1] Laboratoire dÉcologie des Hydrosystèmes Naturels et Anthropisés – Institut National de la Recherche Agronomique : USC1369, Centre National de la Recherche Scientifique : UMR5023, Ecole Nationale des Travaux Publics de l'Etat, Université Claude Bernard Lyon 1 – France

[2] Bases de données sur la Biodiversité, Ecologie, Environnement et Sociétés – Museum National d'Histoire Naturelle, Centre National de la Recherche Scientifique : UMS3468 – France

GOTIT est une base de données relationnelle et une application Web qui permettent d'optimiser la productivité d'un laboratoire, notamment dans les domaines de la conception de projets de recherche, de l'analyse des biais et de la planification d'échantillonnage, de l'identification des espèces, du séquençage de l'ADN, du biobanking, et du transfert des données vers des plateformes mondiales de la biodiversité. GOTIT gère les différentes étapes d'un processus de production de données d'occurrence d'espèces, allant de l'échantillonnage au séquençage, en passant par le stockage de lots d'individus, de lames et d'extraits d'ADN et leur affectation à des hypothèses d'espèces fondées sur la morphologie et l'ADN. L'application prend également en compte les données d'occurrence d'espèces et les métadonnées de séquences d'ADN provenant de sources externes.

GOTIT est conçu pour optimiser les activités de recherche sur la diversité et l'évolution des espèces à forte composante géographique, telles que la biogéographie et la phylogéographie. Une caractéristique clé de GOTIT est sa capacité à assigner de multiples hypothèses d'espèces fondées sur des méthodes de délimitation morphologiques et moléculaires à un même ensemble d'individus. Une autre caractéristique clé est la traçabilité, qui favorise la répétabilité scientifique et le partage des tâches entre utilisateurs.

*Intervenant