Table des matières

Data curation, fisheries and ecosystem-based management: the case study of the Pecheker database

Alexis Martin * [1], Charlotte Chazeau * †

[1] Biologie des Organismes et Ecosystèmes Aquatiques (BOREA) – Muséum National d'Histoire Naturelle (MNHN) – 7, rue Cuvier, CP 32, 75231 Paris Cedex 05, France

The scientific monitoring of the Southern Ocean French fishing industry is based on the use the Pecheker database. Pecheker is dedicated to the digital curation of the data collected on field by scientific observers and which analysis allows the scientists of the Muséum national d'Histoire naturelle institution to provide guidelines and advice for the regulation of the fishing activity, the protection of the fish stocks and the protection of the marine ecosystems. The template of Pecheker has been developed to make the database adapted to the ecosystem-based management concept. Considering the global context of biodiversity erosion, this modern approach of management aims to take account of the environmental background of the fisheries to ensure their sustainable development. Completeness and high quality of the raw data allowing the building of strong predictive models is a key element for an ecosystem-based management database such as Pecheker. Here, we present the development of this database as a case study of fisheries data curation. Considering the success factors we could identify, we propose a discussion about how the community could build a global fisheries information system based on a network of small databases including interoperabilitystandards.

*Intervenant

†Auteur correspondant: charlotte.chazeau@mnhn.fr

Perturbation de réseaux de neurones pour comprendre le positionnement des nucléosomes chez la souris

Maxime Christophe * [1]

[1] Structure et Instabilité des Génomes – CNRS, Institut National de la Santé et de la Recherche Médicale - INSERM, Muséum National d'Histoire Naturelle (MNHN) – France

Les nucléosomes sont les unités structurantes de la chromatine, formant une base essentielle pour l'organisation du génome et la régulation de son activité. Disposés régulièrement le long de l'ADN, ils jouent un rôle crucial dans la compaction et la protection de l'information génétique, ainsi que dans le contrôle de l'accessibilité aux facteurs de transcription. Cependant les mécanismes précis régissant leur positionnement restent en grande partie inconnus.
Ces dernières années, les réseaux de neurones se sont révélés prometteurs pour prédire les sites de positionnement des nucléosomes, en raison de leur capacité à intégrer des données complexes et à identifier des motifs récurrents. Néanmoins, l'une des principales limitations de ces approches réside dans le problème de la "boîte noire". Cela signifie qu'il est souvent difficile de comprendre les règles exactes apprises par ces réseaux, rendant opaque l'interprétation des résultats.
L'approche du "mutasome", qui consiste à perturber ces réseaux pour examiner les variations dans leurs prédictions, est proposée comme solution pour extraire des règles biologiquement et humainement interprétables. En appliquant cette méthode au modèle murin, il devient possible de mieux comprendre les déterminants du positionnement des nucléosomes, ouvrant ainsi la voie à des avancées dans la compréhension des dynamiques chromatiniennes et de leur rôle dans la régulation du génome.

*Intervenant

Kefir-ensemble: Quelles outils pour l'intégration de données de métagénomiques à grande échelle dans le cadre d'un projet de science participative?

Eliott Tempez [1,2], Marie Cariou [*†] [1], Alexandra Joubert[‡] [2], Evelyne Duvernois-Berthet[§] [2,3], Delphine Mazé[¶] [2], Jean-Baptiste Boulé[‖] [2]

[1] Acquisition et Analyse de Données pour l'Histoire naturelle – Muséum National d'Histoire Naturelle (MNHN), CNRS UAR2700 – France
[2] Structure et Instabilité des Génomes – MNHN, Inserm U 1154, CNRS, UMR 7196, MNHN, Sorbonne Universities, Paris, France – France
[3] Reproduction et développement des plantes – Laboratoire Reproduction et Développement des Plantes, Univ Lyon, ENS de Lyon, UCB Lyon 1, CNRS, INRAE, Lyon, F-69342 France – France

Le projet de science participative Kéfir Ensemble1 vise à étudier la stabilité et l'évolution des communautés microbiennes dans les grains de kéfir au cours de multiples cycles de fermentation. Le Kéfir est une boisson fermentée dont le ferment se présente sous la forme de grains translucides composés d'une communauté symbiotique de micro-organismes (levures, bactéries lactiques et bactéries acétiques). 1000 participants et participantes, résidant en France métropolitaine, cultivent actuellement des grains de Kéfirs issus d'une même communauté initiale. Les grains recueillis au bout de 50 et 100 cycles de fermentation seront analysés, notamment par séquençage métagénomique afin d'étudier les variations dans les communautés microbiennes, variations qui seront mises en relations avec les différences géographiques ou de protocole de culture.
Dans cette présentation, nous aborderons les enjeux liés à l'analyse et l'intégration des grandes quantités de données de séquençage produites dans le cadre de ce projet. L'objectif est de permettre la description, la quantification et la comparaison de communautés bactériennes de nombreux échantillons de façon automatisable et reproductible. Un pipeline réalisé sous snakemake permet de répondre à ces objectifs au moyen d'un outil facilement utilisable et produisant des formats de sortie synthétiques et adaptés pour faciliter l'intégration à venir des données issues du projet Kéfir ensemble.

Références

1. https://www.kefirensemble.org/homepage

[*]Intervenant
[†]Auteur correspondant: marie.cariou@mnhn.fr
[‡]Auteur correspondant: alexandra.joubert@mnhn.fr
[§]Auteur correspondant: evelyne.duvernois-berthet@ens-lyon.fr
[¶]Auteur correspondant: delphine.maze@mnhn.fr
[‖]Auteur correspondant: jean-baptiste.boule@mnhn.fr

PLANETOID: deeP LeArNing-based gEnome annoTation Of bIoDiversity

Julien Mozziconacci * [1]

[1] Structure et Instabilité des Génomes – Museum National d'Histoire Naturelle, Institut National de la Santé et de la Recherche Médicale, Institut de Chimie du CNRS, Sorbonne Universite, Centre National de la Recherche Scientifique – France

PLANETOID, un projet financé dans le cadre de la réponse du "Sorbonne Cluster for Artificial Intelligence" (SCAI) à l'appel ClusterIA de l'ANR, se concentre sur l'annotation extensive et précise de divers génomes et métagénomes. L'objectif est de développer des méthodes automatisées et évolutives pour l'annotation des génomes d'organismes non-modèles, en s'appuyant sur les dernières avancées en intelligence artificielle, notamment les modèles de langages. Grâce à ces nouvelles méthodes, le projet PLANETOID, qui démarrera en 2025, vise à découvrir de nouveaux éléments génétiques et mécanismes régulateurs, tout en constituant une base de données accessible et détaillée. Cette ressource sera essentielle pour les recherches futures en génomique, métagénomique, écologie, et évolution.

*Intervenant

Modéliser des systèmes patrilinéaires pour expliquer le goulot d'étranglement génétique du chromosome Y à la fin du Néolithique

Léa Guyon * [1], Jérémy Guez [1,2], Bruno Toupance [1], Evelyne Heyer [1], Raphaëlle Chaix [1]

[1] Éco-Anthropologie – Muséum National d'Histoire Naturelle (MNHN), Centre National de la Recherche Scientifique - CNRS, Université Paris-Cité – France
[2] Laboratoire Interdisciplinaire des Sciences du Numérique – Centre National de la Recherche Scientifique - CNRS, L'Institut National de Recherche en Informatique et e n Automatique (INRIA), Université Paris-Sud - Université Paris-Saclay – France

Il y a environ 5000 ans, la diversité génétique du chromosome Y, transmis de père en fils, subit un fort déclin dans le monde entier, alors que la diversité de l'ADN mitochondrial, marqueur génétique transmis par la mère, ne diminue pas. Nous faisons l'hypothèse que cette chute brutale de la diversité paternelle est due à un changement social majeur associé à la transition Néolithique, marquant le passage de sociétés de chasseurs-cueilleurs à des sociétés d'agriculteurs et d'éleveurs. Des études antérieures ont montré que les systèmes de parenté, qui déterminent à qui les individus sont apparentés dans une population, quels sont les mariages encouragés ou interdits et où les couples s'installent après leur mariage, ont un impact sur la diversité génétique des populations. En particulier la patrilinéarité (appartenance des individus au lignage et au clan de leur père), et la patrilocalité (résidence près des parents du mari), ont pour effet de diminuer la diversité génétique du chromosome Y par rapport à l'ADN mitochondrial. En utilisant le logiciel SLiM, nous avons modélisé des populations humaines présentant des systèmes patrilinéaires et patrilocaux, en calibrant les paramètres du modèle sur la littérature anthropologique, afin de tester l'hypothèse selon laquelle une transition vers un système patrilinéaire aurait pu engendrer une chute importante de la diversité génétique du chromosome Y. En simulant différents scénarii, nous mettons en évidence qu'un changement vers des systèmes patrilinéaires patrilocaux, caractérisés par une importante variance de succès reproducteur entre groupes de filiation et des fissions linéaires (i.e. le long de la lignée paternelle) régulières de ces groupes de filiation, permet de diminuer fortement la diversité génétique du chromosome Y. De plus, nous montrons que les conflits violents entre groupes de filiation ne sont pas nécessaires pour expliquer le goulot d'étrangement de la diversité génétique paternelle.

*Intervenant

Évolution de l'ADN centromérique chez les cercopithèques

Julien Pichon *† [1], Jensen Axel , Katerina Guschanski , Loic Ponger‡ , Christophe Escudé§

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

Bien que la fonction du centromère soit très conservée, l'ADN qui le compose montre une grande diversité entre les espèces. Chez les primates, l'ADN centromérique le plus abondant est appelé ADN alpha-satellite (AS), composé de monomères répétés en tandem de 171bp. De nombreuses études chez l'humain, ainsi que les assemblages T2T récents, ont permis de caractériser différentes familles d'AS ainsi que l'organisation de ces familles le long des chromosomes. Les monomères d'AS s'étendent bien au delà des centromères, formant des couches où les monomères d'une même couche vont montrer une similarité plus forte que les monomères d'une couche différente. Cette similarité intra-couche diminue au fur et à mesure que l'on se dirige vers les bras des chromosomes.
Pour comprendre l'évolution des AS, nous travaillons sur les cercopithèques, un clade proche de celui des macaques, et où 45 espèces ont divergé en moins de 10 millions d'années. Leur grande quantité d'AS dans le génome ainsi que leurs nombreuses fissions chromosomiques (48 à 72 chromosomes) en font un groupe d'intérêt pour comprendre l'évolution des séquences centromériques. Grâce à une collaboration avec Katja Guschanski de l'Université d'Edimbourg, nous avons pu récemment séquencer en Pacbio HiFi 8 espèces provenant de la collection RBCell du Museum, offrant des reads de haute qualité avec une longueur moyenne d'environ 17kbp. Chaque read couvre ainsi une centaine de monomères, permettant d'étudier leur organisation linéaire.
Je présenterai ici les différents outils bioinformatiques que nous avons développé afin d'étudier l'évolution des AS directement sur les reads. Cela inclut la détection et la classification des monomères d'AS à l'aide de méthode non-supervisée ou encore l'étude de l'organisation linéaire des familles via des calculs de similarité. Je montrerai les résultats obtenus pour une espèce de cercopithèque, mettant en évidence l'existence de plusieurs familles d'AS avec des histoires évolutives distinctes.

*Intervenant
†Auteur correspondant: julien.pichon@mnhn.fr
‡Auteur correspondant: loic.ponger@mnhn.fr
§Auteur correspondant: christophe.escude@mnhn.fr

Intelligence artificielle et annotation : un outil essentiel pour les experts

Tristan Tchilinguirian * [1]

[1] Muséum National dHistoire Naturelle – Muséum National d'Histoire Naturelle (MNHN) – France

Dans le cadre des systèmes d'intelligence artificielle (IA), on parle souvent de "machines qui apprennent à partir de données". Cependant, la qualité de ces données est primordiale pour la performance des algorithmes d'apprentissage
Pour cela, il est indispensable d'avoir des données annotées par des spécialistes qui connaissent bien les objets à identifier.
C'est là que l'importance d'un **outil d'annotation performant** entre en jeu, un outil accessible, efficace et surtout **adapté aux besoins des experts**.
CVAT : un outil libre, déployable et flexible

L'outil **CVAT** est une solution open source conçue spécifiquement pour l'annotation d'images et de vidéos. Ce logiciel offre plusieurs avantages :

- Il peut être déployé **localement ou sur un serveur interne**.
- Les utilisateurs peuvent également accéder à l'outil à distance, facilitant ainsi la collaboration avec des agents extérieurs, tout en restant sécurisé.
- **CVAT** est pensé pour être **user-friendly**, ce qui permet aux non-informaticiens de s'en servir efficacement pour annoter les images.

L'annotation facilitée par l'IA : CVAT et ses outils avancés
En plus de son interface conviviale, **CVAT** propose des outils puissants d'annotation semi-automatique et automatique grâce à l'intégration de modèles d'intelligence artificielle.
Par exemple, des outils comme **Deep Extreme Cut** ou **Mask R-CNN** permettent une annotation plus rapide en détectant automatiquement les objets à partir d'un petit ensemble de points
(promesse de multiplier la vitesse d'annotation par 10)
Ces modèles sont adaptables à différents types de données (formes complexes, objets 3D, etc.) et offrent des résultats précis, tout en restant modulables par les utilisateurs.

*Intervenant

Analyse bioinformatique de données de transcriptomique spatiale : Exploration de l'effet combiné de l'exposition aux perturbateurs endocriniens et au régime gras et sucré sur la signature transcriptomique des différentes noyaux de l'hypothalamus chez la souris

Pilar Rodríguez * [1,2], Anni Herranen [3], Evelyne Duvernois-Berthet [4,5], Justine Fredoc-Louison [3], Isabelle Seugnet [3], Amina Mahi-Moussa [3], Marie-Stéphanie Clerget-Froidevaux [6]

[1] Département Adaptations du vivant – Museum National d'Histoire Naturelle - MNHN (FRANCE) – France
[2] Acquisition et Analyse de Données pour l'Histoire naturelle – UMS 2700 (2AD) – France
[3] Physiologie moléculaire et adaptation – Muséum National d'Histoire Naturelle (MNHN), CNRS – France
[4] Structure et Instabilité des Génomes – MNHN, Inserm U 1154, CNRS, UMR 7196, MNHN, Sorbonne Universities, Paris, France – France
[5] Reproduction et développement des plantes – Laboratoire Reproduction et Développement des Plantes, Univ Lyon, ENS de Lyon, UCB Lyon 1, CNRS, INRAE, Lyon, F-69342 France – France
[6] Physiologie moléculaire et adaptation – Muséum National d'Histoire Naturelle (MNHN), CNRS – France

Le haut taux d'obésité n'est pas seulement le résultat d'un mode de vie sédentaire et d'un régime enrichi en graisse et en sucre mais aussi de l'exposition environnementale aux perturbateurs endocriniens (PE). Certains perturbateurs appelés obésogènes ciblent des réseaux centraux régulant le comportement alimentaire et la consommation d'énergie, particulièrement dans l'hypothalamus. En effet, cette zone du cerveau est le régulateur central de plusieurs fonctions incluant l'activité des hormones thyroïdiennes et le métabolisme. L'hypothalamus est composé de plusieurs noyaux, chacun ayant un rôle unique dans ces fonctions. Les organismes sont vulnérables aux perturbateurs endocriniens pendant la période périnatale lorsque la signalisation hormonale dirige la configuration des axes endocrinien et métabolique. Dans ce contexte, cette étude vise à explorer l'effet combiné de l'exposition périnatale au perturbateur tétrabromobisphénol A (TBBPA), qui interrompt la signalisation de l'hormone thyroïdienne, et du régime enrichi en graisse et en sucre (HFHS) à l'âge adulte sur le transcriptome de l'hypothalamus chez deux souches de souris. Afin de mesurer l'activité transcriptionnelle dans les différentes noyaux de l'hypothalamus en réponse aux différents traitements, l'étude a utilisé la transcriptomique spatiale. Cette technologie permet d'identifier la distribution spatiale des gènes exprimés dans le tissu. La classification automatique des groupes de types cellulaires différents par rapport à leur profil d'expression est une étape cruciale dans l'analyse bioinformatique de ce type de données. Étant donné que le résultat de cette étape dépend du type de tissu, sa taille et les algorithmes

*Intervenant

d'analyse utilisés, sa mise en oeuvre représente un défi. Cette présentation sera focalisée sur les différentes stratégies abordées pour améliorer la classification automatique des noyaux de l'hypothalamus et l'identification de leurs signatures transcriptionnelles.