# De l'INRA au MNHN en passant par le CNRS et INRIA, retour d'expérience sur 10 ans liés à la bioinformatique

Yvan Le Bras*[1]

[1]MNHN CESCO Station de Biologie Marine de Concarneau (MNHN) – Muséum National d'Histoire Naturelle (MNHN) – France

**Résumé**

Comme beaucoup j'ai découvert le terme " bioinformatique " quand mes usages de biologistes ne me permettaient pas de passer à l'échelle, ne pouvant utiliser Excel... Depuis mes débuts en génétique des populations, je n'ai alors eu cesse de plonger toujours plus profondément dans cette infâme environnement hostile que l'on nomme informatique, jusqu'à atteindre les fosses abyssales de l'informatique scientifique, du " HPC " puis enfin du cloud computing et des microservices. Peut-être par chance, je ne m'y suis pas noyé, et n'ai eu cesse d'en ressortir l'essentiel, le positif, qui peut aujourd'hui être résumé dans ce que porte l'approche " FAIR ". Expression entrée dans le langage courant des infrastructures de recherche, la " FAIRisation " de la recherche se veut rendre la donnée scientifique plus Trouvable, Accessible, Intéropérable et Réutilisable (" FAIR : Findable, Accessible, Interoperable, Reusable ") et induit le recours à de nombreux outils et services de l'informatique scientifique et à ce mouvement fou ;) qu'est celui de la science ouverte. De mon parcours entre usage et mise à disposition de services en bioinformatique, je propose de présenter une vision assez personnelle de la bionformatique à travers les expériences vécues et notamment, mais pas que, mon expérience autour de ce magnifique projet open-source Galaxy.

*Intervenant

# Génomique comparative réduite du mammouth laineux : comment la bioinformatique peut conduire à modeler le design expérimental.

Laetitia Aznar-Cormano*[1], Jawad Abdelkrim[2], and Régis Debruyne[3]

[1]Centre de recherche sur la Paléobiodiversité et les Paléoenvironnements – Museum National d'Histoire Naturelle, Sorbonne Universite, Centre National de la Recherche Scientifique : UMR7207 – France
[2]DGD REVE - UMS 2700 - Service de Systématique Moléculaire – Musée National d'Histoire Naturelle - MNHN (France) – France
[3]DGD REVE – Muséum National d'Histoire Naturelle (MNHN) – France

**Résumé**

Le Restriction-Associated DNA sequencing (RADseq) est une méthode désormais très populaire en génomique comparative réduite pour les taxons non modèles. Néanmoins, son application est délicate pour du matériel génétique dégradé, fréquemment rencontré dans les collections de musée. Des approches expérimentales ont été proposées pour contourner cette difficulté en intégrant une approche de capture ciblée de la totalité (hyRAD) ou d'une partie (Rapture, RADcap) des catalogues de marqueurs RAD. Pour ces dernières, une étape supplémentaire de sélection de sondes de capture doit être implémentée.
Cette communication présente comment nous avons choisi d'optimiser un protocole de capture par sondes nucléaires pour des mammouths laineux (*Mammuthus primigenius*). En particulier, j'aborderai deux aspects clés : l'analyse in-silico du génome complet de *Loxodonta africana* à des fins d'optimisation du protocole de RADseq sur un échantillon d'éléphants modernes, et l'analyse des données issues de ce séquençage RADseq afin de définir un catalogue de 20,000 sondes de capture distribuées à travers le génome nucléaire du mammouth laineux.

*Intervenant

# GBIF – Global Biodiversity Information Facility et réseau des portails nationaux " Living Atlases "

Anne-Sophie Archambeau*[1], Marie-Elise Lecoq[2], Sophie Pamerlon[2], Fabien Cavière[2], Eric Chenin[3], and Régine Vignes-Lebbe[4]

[1]IRD/MNHN/UMS PatriNat/GBIF France – Muséum National d'Histoire Naturelle (MNHN) : UMSPatriNat – MNHN CP48 43 rue Buffon 75005 Paris, France

[2]Muséum national d'histoire naturelle (GBIF France/MNHN) – Ministère de l'Enseignement Supérieur et de la Recherche, Muséum National d'Histoire Naturelle (MNHN) – CP 48 - 43 rue Buffon 75005 Paris, France

[3]Institut de Recherche pour le Développement (IRD) – Institut de Recherche pour le Développement – 32, avenue Henri Varagnat, 93143 Bondy cedex, France

[4]Institut de Systématique, Evolution, Biodiversité (ISYEB) – Muséum National d'Histoire Naturelle (MNHN), CNRS : UMR7205, Université Pierre et Marie Curie (UPMC) - Paris VI, EPHE, Université Pierre et Marie Curie [UPMC] - Paris VI – France

**Résumé**

Le point nodal français du GBIF est intégré dans l'UMS-PatriNat (AFB-CNRS-MNHN) et travaille de concert avec SINP/INPN et PNDB (Pôle National de Données de Biodiversité).
Le GBIF France contribue à la connection des données françaises au GBIF international, en mettant à disposition des communautés scientifiques et politiques françaises les outils, standards et services créés par le GBIF. A ce jour, le portail GBIF (www.gbif.org) donne accès à plus d'un milliard d'occurences provenant de collections ou d'observations téléchargeables librement.

Cet exposé présentera les dernières fonctionalités du portail : mise en place de licences, metrics, DOI sur tous les jeux de données et sur tous les téléchargements effectués qui facilitent la citation et permettent un meilleur suivi de l'utilisation des données, ainsi que le lien direct vers les publications associées aux jeux de données.
Nous présenterons également la communauté des " Living Atlases ", basés sur les modules développés par Atlas of Living Australia, (spatial, portail données et métadonnées, ALA4R...) et dont le but est d'aider toute institution qui le souhaite, à créer leur portail national ou thématique. Actuellement 13 portails sont en production dans le monde (dont portail GBIF France) et d'autres sont en cours de réalisation (dont requeteur SINP).

*Intervenant

# Exploration de la matière noire omique à partir de méthodes de réseaux

Lucie Bittner*[1,2]

[1]Institut de Biologie Paris Seine, Université Pierre et Marie Curie (IBPS, UPMC) – Université Pierre et Marie Curie [UPMC] - Paris VI – Université Pierre et Marie Curie Institut de Biologie Paris Seine 7-9 Quai Saint Bernard 75252 Paris cedex 05, France

[2]Institut de Systématique, Evolution, Biodiversité UMR 7205 – Centre National de la Recherche Scientifique, Ecole Pratique des Hautes Etudes, Sorbonne Universite, Museum National d'Histoire Naturelle – France

**Résumé**

Notre compréhension du monde microbien vit un changement de paradigme. La métagénomique et la métatranscriptomique offrent des mesures sans précédent de la biodiversité taxonomique et fonctionnelle des communautés. Cependant, les outils d'inférence actuels reposant principalement sur des noms d'espèces ou des noms de fonctions, une part importante des séquences (méta-)omiques est ignorée. Les progrès des méthodes de culture et l'augmentation du nombre d'organismes modèles aident à réduire la proportion de séquences inconnues, mais ces solutions coûteuses sont difficiles à mettre en œuvre rapidement. À l'heure actuelle, il existe des méthodes bioinformatiques capables d'exploiter l'énorme quantité de séquences connues et inconnues, et qui permettent ainsi de dépasser notre vision encore incomplète des communautés. A l'occasion de ces rencontres, je propose d'exposer les méthodes utilisées dans mon équipe afin d'explorer la matière noire omique microbienne. Je présenterai notamment des travaux utilisant des réseaux de similarité de séquences et des réseaux de cooccurrence. Ces réseaux permettent d'étudier sans *a priori* les processus adaptatifs et évolutifs qui façonnent la diversité taxonomique et fonctionnelle des organismes non modèles dans l'environnement, et offrent de nouvelles perspectives pour stimuler et enrichir les modèles décrivant la dynamique des populations ou encore les cycles biogéochimiques.

*Intervenant

# La bio-informatique appliquée dans le cadre du Programme d'Observation Écosystémique des Pêcheries Australes

Patrice Pruvost*[1], Alexis Martin*[2], Clara Péron , Charlotte Chazeau , Nicolas Gasco , and Guy Duhamel

[1]Biologie des Organismes et Ecosystèmes Aquatiques (BOREA) – Sorbonne Universite : UM95, Museum National d'Histoire Naturelle, Centre National de la Recherche Scientifique, Université de Caen Normandie : UMR7208, Institut de Recherche pour le Développement – 7, rue Cuvier, CP 32, 75231 Paris Cedex 05, France

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

**Résumé**

Le MNHN assure le suivi scientifique des pêcheries australes françaises depuis 1978 et fournit des avis scientifiques pour l'encadrement de la pêche aux administrations gestionnaires.
Ces avis, longtemps fondés sur d'évaluation halieutique, reposent aujourd'hui sur une approche écosystémique intégrant différents travaux de modélisation et des études sur l'impact de la pêche sur le milieu et les habitats naturels.

La modélisation des biomasses des espèces ciblées (poissons, langoustes) et des impacts sur les autres espèces sont déterminantes pour établir des recommandations sur les quotas de pêche. Une seconde approche repose sur la construction de modèles spatiaux pour mettre en évidence les patrons de distribution des différentes espèces et caractériser leur niche écologique. L'étude des assemblages d'espèces permet de cartographier les écosystèmes marins benthiques nécessaires à la protection de la biodiversité.

Notre équipe développe un système d'information complexe qui intègre l'ensemble des données collectées à bord des navires commerciaux et scientifiques. Ces données incluent des observations d'interaction et d'impacts sur le milieu et les habitats. Les données sont réparties dans quatre bases du Muséum (Pecheker, BasExp, GICIM, InvMar).
Ces travaux de bio-informatique nécessitent des ressources importantes en matériels, moyens de calcul et des compétences scientifiques des ingénieurs et des chercheurs impliqués.

*Intervenant

# Sélection divergente de caractères polygéniques : simulation et détection

Léa Bouteille* and Frédéric Austerlitz[1]

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

**Résumé**

Les populations d'une même espèce vivent souvent dans des environnements différents, les optima phénotypiques diffèrent pour de nombreux caractères entre ces populations. Ces caractères sont donc sous sélection divergente. Ils peuvent être monogéniques ou polygéniques, mais les méthodes de détection de gènes sous sélection les plus couramment utilisées ont été développées dans le cadre de la sélection monogénique. En utilisant le programme quantiNEMO, nous avons simulé un ensemble de populations reliées par des flux de gènes et soumises à une sélection divergente sur un caractère polygénique. Nous avons montré que dans beaucoup de cas, les fréquences alléliques des locus impliqués dans le caractère polygénique ne différaient pas beaucoup d'une population à l'autre. Or les méthodes de détection recherchent les locus présentant un excès de différenciation parmi un ensemble de locus supposés neutres. Nous avons comparé quatre méthodes de détection (FDIST, BayeScan, OutFLANK et PCAdapt). Dans toutes les conditions, aucun des programmes n'a été très efficace pour détecter les gènes impliqués dans les caractères polygéniques sous sélection, mais nous avons montré entre autres que certains sont plus conservateurs que d'autres en fonction des conditions.

*Intervenant

# Phénotypes métaboliques et modélisations canoniques

Alain Paris [1]

[1] MNHN – MCAM

**Résumé**

La métabolomique réalise le phénotypage des organismes au plus près de la réalité physiologique exprimée. Elle s'appuie sur le tryptique physiologie-chimie analytique-statistique. L'arsenal bioinformatique s'adresse à ces trois piliers sous des configurations qui leur sont adaptées. En particulier, pour l'exploitation des données produites par les générateurs de variables que sont la spectrométrie de masse et la résonance magnétique nucléaire, elles couvrent en amont le prétraitement des données, directement en lien avec la question physiologique étudiée la modélisation statistique multivariée des données, enfin après la détection des biomarqueurs métaboliques putatifs la recherche sous différentes formes d'éléments de corroboration de leur structure chimique. Quelques exemples illustreront la modélisation canonique régularisée des données mais aussi une modélisation canonique particulière utilisée pour identifier au plan structural de nouveaux métabolites.

# La bioinformatique à l'IMPMC : de la structure aux génomes

Isabelle Callebaut*[1], Jacques Chomilier , Bruno Collinet , Manuela Dezi , Elodie Duprat , Stéphanie Finet , Jean-Michel Guignier , Marie-Anne Hervé Du Penhoat , Slavica Jonic , Mélanie Poinsot , Feriel Skouri-Panet , Dirk Stratmann , and Catherine Vénien-Bryan

[1]Institut de minéralogie, de physique des matériaux et de cosmochimie – Museum National d'Histoire Naturelle, Institut de recherche pour le développement [IRD] : UR206, Sorbonne Universite : UM120, Centre National de la Recherche Scientifique : UMR7590 – France

**Résumé**

L'équipe Bioinformatique et Biophysique de l'IMPMC étudie les bases moléculaires des fonctions et de l'évolution des macromolécules biologiques grâce à un large spectre d'outils théoriques et expérimentaux.
Différentes approches sont développées pour étudier les structures, fonctions et dynamiques des biomolécules à différentes échelles, depuis la compréhension des évènements moléculaires survenant lors de leur irradiation jusqu'à l'analyse de la variabilité conformationnelle de complexes par analyse d'images issues de la cryo-microscopie électronique. L'équipe a recours à la modélisation et/ou à des simulations de dynamique moléculaire (incluant des techniques avancées telles que la métadynamique) pour étudier les fonctions des protéines et leurs interactions avec des partenaires (*e.g.* protéines, peptides cycliques, petites molécules, acides nucléiques), en particulier dans le but d'estimer l'impact de mutations et de développer des thérapies rationnelles.
Par ailleurs, nous nous intéressons également au répertoire structural et fonctionnel des génomes, avec une emphase particulière sur les éléments conservés des repliements de protéines, la considération de ces éléments pour décoder le "dark proteome" (séquences non encore annotées) et sur l'identification d'évènements évolutifs de diversification fonctionnelle. Ces développements sont en particulier appliqués à l'analyse de protéines d'intérêt médical (prédiction de l'impact de mutations) et environnemental (études des mécanismes moléculaires de biominéralisation).

*Intervenant

# Equipe ADN Répété, Chromatine, Evolution (ARChE)

Christophe Escudé*[1]

[1]Structure et Instabilité des Génomes, MNHN – CNRS : UMR7196 – France

**Résumé**

Des séquences d'ADN répétées en tandem sont présentes en grande quantité au niveau des régions centromèriques des chromosomes chez quasiment tous les organismes eucaryotes. Ce composant des génomes, relativement peu étudié, est l'objet de mécanismes d'évolution spécifiques. Par ailleurs, la distribution particulière de ces séquences dans le noyau pourrait jouer un rôle dans la mise en place de l'organisation tridimentionnelle du génome et contribuer à des processus de régulation de l'expression génique. L'équipe ARChE vise à mieux comprendre l'évolution de ces séquences ainsi que leur rôle dans le fonctionnement de la cellule. Leur composition chez différentes espèces de cercopithèques a été abordée par des approches de séquençage à haut-débit. Des méthodes d'analyse bioinformatique adaptées à l'étude de ces séquences, développées au sein de l'équipe, ont permis d'étudier les différences inter-espèces et de proposer un scenario évolutif. Grâce à un outil d'analyse d'image en 3 dimensions spécifiquement dédié à l'étude de l'organisation du génome dans le noyau, il a été possible, pour la première fois, de décrire de façon systématique et quantitative l'organisation nucléaire des régions centromériques dans le noyau de lymphoblastes humains, par une approche d'imagerie à haut-débit. Ces différents développements seront présentés ainsi que leur utilité pour la communauté.

*Intervenant

# Ecouter dans le noir: depuis le traitement du signal jusqu'aux modélisations à large échelle, un exemple d'utilisation intensive de l'informatique en écologie.

Jean-Francois Julien*[1], Yves Bas , and Grégoire Loïs

[1]Centre d'écologie et de sciences de la conservation, CESCO, UMR7204 – Muséum National d'Histoire Naturelle (MNHN) – France

**Résumé**

L'écholocation que pratiquent les chauves-souris de façon quasi continue nous offre une opportunité unique de suivre leur activité nocturne. L'apparition d'enregistreurs autonomes a permis d'automatiser ce suivi, depuis l'acquisition des sons sur le terrain jusqu'à leur interprétation sur le plan écologique. L'intégration de ce processus dans un programme de science participative, Vigie Chiro, nous fournit désormais une masse importante de données, quelques millions de fichiers de son par an, représentant une à quelques dizaines de téraoctets.
Chaque étape du traitement recourt à des techniques informatiques bien spécifiques:
1) détection et mesures des signaux, reposant sur des transformées de Fourier. 2) identification des espèces par des méthodes d'apprentissage machine, Random Forest en l'occurence. 3) gestion des données dans une base mongoDB, 4) analyse, interprétation et modélisation surtout basées sur l'emploi de modèles linéaires généralisés.
Le centre de calcul de l'IN2P3 (cc.in2p3.fr) représente notre ressource principale en termes de stockage et de calcul.

*Intervenant

# Exemples choisis d'utilisation d'outils de bio-informatique pour le traitement de données omiques, dans le cadre de l'étude des effets des proliférations de cyanobactéries chez les poissons

Benjamin Marie*[1,2]

[1]Muséum national d'histoire naturelle (MNHN) – Ministère de l'Ecologie, du Développement Durable et de l'Energie, Ministère de l'Enseignement Supérieur et de la Recherche, Muséum National d'Histoire Naturelle (MNHN) – 57, rue Cuvier - 75231 Paris Cedex 05, France
[2]Musée National d'Histoire Naturelle - MNHN (France) – Musée National d'Histoire Naturelle - MNHN (France) – 12 rue Buffon 75005 Paris, France

**Résumé**

Les prolifrations de cyanobactéries sont suceptibles de perturber la biologie des organismes aquatiques qui y sont exposés, notamment à travers la production de nombreux métabolites bio-actifs. Certains sont des toxines notoires, on les appelle alors cyanotoxines, et ils sont susceptibles d'induire des effets délétères chez les vertébrés.
A travers une série d'exemples choisis à partir des travaux de recherche menés ces dernières années, nous illustrerons comment différentes approches omiques, utilisant des outils de bioinformatiques spécifiques, permettent d'avancer dans cette thématique.

- Annotation des génomes de cyanobactéries à l'aide de la plateforme MicroScope.

- Intégration qualitative de données de protéomique et de transcriptomique chez des poissons exposés aux toxiques avec Ingenuity Pathway Analysis.
- Projet d'étude couplé des effets au niveau du microbiome et du métabolome associé grace aux outils QIIME 2, MixMC et Diablo (package MixOmics).

*Intervenant

# Données de biodiversité : les outils au service de la chaine de l'information

Thomas Milon* , Laurent Poncet , Solène Robert[1], Remy Jomier , Judith Panijel , Mathieu Clair , and Frédérique Vest

[1]Muséum national d'histoire naturelle (MNHN) – Ministère de l'Ecologie, du Développement Durable et de l'Energie, Ministère de l'Enseignement Supérieur et de la Recherche – 57, rue Cuvier - 75231 Paris Cedex 05, France

**Résumé**

L'UMS PatriNat (AFB, CNRS, MNHN), basée au MNHN, regroupe le SINP/INPN, le PNDB (Pôle National de Données de Biodiversité) et le point nodal français du GBIF.
Un des principaux objectifs de l'UMS est de donner accés le plus largement possible aux données de biodiversité au niveau national. Cet accès est possible uniquement si toute la chaine de l'information, depuis la collecte de la donnée, fonctionne. Ceci se traduit par la mise en place de flux de données standardisées (Darwin Core, EML, SINP...), d'outils et services d'acquisition, de gestion, de traitement, de valorisation et de diffusion de données, autant pour le grand public (ex : INPN espèce) que pour les professionnels de la donnée biodiversité (ex : API).

Nous présenterons ici l'organisation de cette chaine de l'information, les outils et services permettant qu'elle fonctionne, ainsi que les biais qu'elle peut connaitre, impactant directement l'utilisation finale de la donnée.

*Intervenant

# Plateforme Xper3 et collections

Régine Vignes-Lebbe*[1], Gilles Bertin*[2], Thomas Bottini*[2], Sylvain Bouquin*[1], Lisa Chupin*[3], Julien Husson*[4], Eva Perez*[4], Marc Pignal*[4], Rémy Portier*[5], and Manuel Zacklad*[2]

[1]Sorbonne Université, UMR Isyeb – Sorbonne Université UPMC Paris VI – France
[2]CNAM – Laboratoire Dicen-Idf EA 7939 – France
[3]Université Paris-Descartes, Laboratoire Dicen-IDF – Université Paris Descartes - Paris 5 – France
[4]MNHN – Muséum National d'Histoire Naturelle (MNHN) – France
[5]CNRS – UMS 3468 (BBEES) – France

**Résumé**

Xper3 (www.xper3.fr) est une plateforme collaborative facilitant l'informatisation de descriptions structurées (spécimens ou taxons), formalisées selon un modèle descriptif commun comparable à une ontologie de domaine. Xper3 est accompagné de webservices pour l'identification. Au MNHN ce logiciel est utilisé pour des enseignements, dans des programmes de sciences participatives et pour la recherche.
La liaison entre Xper3 et le bureau virtuel de consultation des collections (collaboratoire de l'infrastructure e-Recolnat ANR-11-INBS-0004) est en cours de développement. En effet l'avènement de programmes de numérisation crée un vaste corpus d'images associées aux métadonnées de leurs occurrences et nécessite de nouveaux outils pour travailler directement sur les images numériques des spécimens. Le programme Annotate est un logiciel d'annotation graphique d'images développé conjointement entre le CNAM et le MNHN. Initialement conçu comme un logiciel d'annotation en texte libre, son couplage avec le logiciel Xper3 devrait apporter des référentiels et des vocabulaires contrôlés permettant de sémantiser les annotations.
L'indexation humaine via Annotate ou les Herbonautes et leur généralisation par taxon, permettra de disposer d'un corpus d'images annotées suffisant pour procéder à l'entrainement d'un réseau de neurones. Des premiers tests encourageants ont été réalisés avec des réseaux de neurones convolutionnels pour reconnaître quelques caractères foliaires.

*Intervenant

# Deep learning for genomics

Julien Mozziconacci*[1]

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

**Résumé**

I will introduce deep neural networks and show how they could change practices in functional and comparative genomics.

*Intervenant

# Cribles génomiques et domestications de gènes

Céline Cattelin[1], Pierre Brézellec[1,2], and Sophie Pasek*[2,3]

[1]UVSQ – Université de Versailles Saint-Quentin-en-Yvelines – France
[2]ISYEB – Muséum National d'Histoire Naturelle (MNHN) – France
[3]UPMC – Université Pierre et Marie Curie - Paris 6 – France

**Résumé**

Lorsqu'on cherche à identifier les gènes associés à un trait, on recourt fréquemment à l'utilisation de cribles génomiques. Ainsi, étant donnés des organismes qui partagent le trait que l'on souhaite étudier (groupe IN) et d'autres qui ne présentent pas cette caractéristique (groupe OUT), le crible génomique va identifier les gènes communs aux génomes du groupe IN et absents des génomes du groupe OUT. Les gènes ainsi criblés sont susceptibles d'être liés au trait considéré et peuvent donner lieu à d'autres investigations.
Dans certains cas, chez les bactéries en particulier où les transferts horizontaux sont fréquents, des gènes portés par des phages (ou des plasmides) circulent dans les génomes et peuvent être domestiqués. Une fois domestiqués, ces gènes font partie intégrante du génome des bactéries et peuvent être associés à un trait particulier.
Comme nous l'avons montré dans de précédents travaux, des gènes impliqués dans le cycle cellulaire de certaines bactéries ont été domestiqués. Nous montrerons que les cribles classiques ne sont pas en mesure de les identifier. Nous proposerons une solution pour remédier à ce problème et illustrerons sa pertinence sur ces exemples.

*Intervenant