Recyclage des données de la recherche en biologie : vers une bioinformatique écologique ?

Pierre Poulain

Habilitation à diriger des travaux de recherche (HDR)

« Recyclage des données de la recherche en biologie : vers une bioinformatique écologique ? »

Pierre Poulain

Maître de conférences à l'Université Paris Cité

Équipe « Mitochondries, métaux et stress oxydatif »
dirigée par Jean-Michel Camardro
Institut Jacques Monod
UMR7592 CNRS et Université Paris Cité
15 rue Hélène Brion, 75013 Paris

Présentée et soutenue publiquement le 23 mars 2022 devant le jury composé de :

M. Olivier Taboureau, professeur Université de Paris	Président
Mme Sarah Cohen-Boulakia, professeur Université Paris-Saclay	Rapportrice
M. Matthieu Montes, professeure Conservatoire National des Arts et Métiers	Rapporteur
Mme Julie Thompson, directrice de recherche CNRS	Examinatrice
M. Laurent Gatto, professeur Université Catholique de Louvain	Examinateur

Université Paris Cité

Sommaire

L'organisation de ce document est telle que demandée par l'Université de Paris.

Introduction

Chimiste de formation (2003), physicien par hasard (2006), devenu ensuite bioinformaticien, je suis maître de conférences à l'Université de Paris depuis 2007. Mon parcours universitaire et plus généralement professionnel traduit mon goût prononcé pour les défis aux interfaces, humaines comme scientifiques, et le travail en équipe. Depuis plusieurs années, je manifeste un fort intérêt pour la résolution de problèmes en biologie par des moyens informatiques.

Mon expérience de 4 ans (2012–2016) en République du Congo a été une véritable source de transformation, tant par la réflexion sur mes pratiques pédagogiques que la découverte de nouveaux centres d'intérêt scientifiques.

De retour en France en 2016, avec plein d'idées en tête et une place à trouver, j'ai mené cette transformation tant scientifique que pédagogique. Après 2 ans de formation à la pédagogie (innovante ?) et la production d'un bilan réflexif, c'est finalement mon « HDR de la pédagogie » qui a abouti en 2019 avec le certificat de pédagogie CertifiENS. Mais si je suis enseignant, je suis aussi chercheur, et les questions scientifiques autour des données en biologie m'enthousiasment. La maturation de ces questions m'a conduit à écrire ce manuscrit en vue d'obtenir mon HDR.

Partie 1

Curriculum vitae

Pierre Poulain

42 ans, marié, 2 enfants

Équipe « Mitochondries, métaux et stress oxydatif »
Institut Jacques Monod
UMR7592 CNRS et Université de Paris
15 rue Hélène Brion
75013 Paris

☎ +33 (0)1 57 27 80 28
http://cupnet.net
✉ pierre.poulain@u-paris.fr

Maître de conférences en bioinformatique à l'Université de Paris (ex-Paris Diderot) depuis 2007. Publication de 26 articles dans des revues internationales à comité de lecture (7 comme 1^{er} auteur, 5 comme dernier auteur). Présentation de 17 posters dans des conférences nationales et internationales. Développement de 3 logiciels (`PTools`, `PBxplore`, `AutoclassWrapper`), d'une application web (`Pixel`) et de 2 applications mobiles (`EduPalu` et `DensiPara`), tous sous licence open source. Encadrement scientifique de 2 étudiants en thèse, de 4 étudiants de Master 2, de 4 étudiants de Master 1, de 3 étudiants de Licence 3 et de 8 développeurs.

1.1 Expériences

depuis 2017 **Maître de conférences**, Université de Paris, Paris, France.
Équipe « Mitochondries, métaux et stress oxydatif »
Institut Jacques Monod, UMR7592 CNRS et Université de Paris.

Développement de méthodes et d'outils pour l'analyse de données issues d'expériences de protéomique à haut débit, notamment en spectrométrie de masse.

2014 → 2016 **Chercheur**, Fondation Congolaise pour la Recherche Médicale (FCRM), Brazzaville, République du Congo.

Analyse de données cliniques. Formation aux bonnes pratiques de recherche scientifique. Développement de solutions numériques pour combattre le paludisme avec la plateforme congolaise de développement *open-source* Fongwama. Production de 2 applications mobiles pour lutter contre le paludisme (`EduPalu` et `DensiPara`), disponibles sur GooglePlay. `EduPalu` est arrivée 2^{e} du RFI App Chalenge Afrique 2016.

2014 → 2016 **Analyste de données géospaciales**, Total Exploration & Production Congo, Pointe-Noire, République du Congo.

Développement en Python d'une solution d'intégration de données et de *reporting* pour le système d'information géographique. Ce projet a remporté le Baril d'Or 2015 à Total E&P Congo.
Support pour la gestion de données. Animation d'une formation interne à Python.

2013 → 2014 **Chargé de projets informatiques**, Total Exploration & Production Congo, Pointe-Noire, République du Congo.

Revamping et organisation des salles serveurs de l'entreprise. Gestion documentaire. PRA.

2007 → 2012 **Maître de conférences**, Université Paris Diderot, Paris, France.
Équipe « Dynamique des Structures et Interactions des Macromolécules Biologiques (DSIMB) »
Laboratoire Inserm U665 « Protéines de La Membrane Erythrocytaire et Homologues Non-Erythroides »
Institut National de la Transfusion Sanguine (INTS)

Activités de recherche en bioinformatique structurale. Étude des polymorphismes de l'intégrine humaine αIIbβ3 par simulations numériques (dynamiques moléculaires). Mise en place d'une base de données rassemblant des informations cliniques et des données issues de cytométrie en flux – application à la drépanocytose.

2006 → 2007 **Post-doctorant**, CNRS, Paris, France.
Laboratoire de Biochimie Théorique (LBT)
Institut de Biologie Physico-Chimique (IBPC)

Développement d'un modèle gros grain pour l'amarrage protéine/ADN.

2003 → 2006 **Doctorant**, Université Claude Bernard Lyon 1, Lyon, France.
Laboratoire de spectrométrie ionique et moléculaire

Titre de la thèse : « Structure et dynamique de protéines isolées : approches statistiques »

1.2 Diplômes

2006 **Doctorat de physique**, Université Claude Bernard Lyon 1, Lyon, France.

2003 **MSc in Cheminformatics**, UMIST, Manchester, Royaume-Uni.

2003 **Diplôme d'ingénieur chimiste**, École Nationale Supérieure de Chimie de Montpellier (ENSCM), Montpellier, France.

1.3 Responsabilités collectives

depuis 2021 **Membre élu du conseil d'enseignement** de l'UFR Sciences du Vivant.

depuis 2021 **Ambassadeur Software Heritage**. Communication et développement des bonnes pratiques pour l'archivage des codes sources logiciels dans la communauté bioinformatique.

depuis 2019 **Membre du comité technique de la plateforme iPOP-UP** (*Integrative Platform for Omics Projects at Université de Paris*).

depuis 2018 **Membre du comité de pilotage de la plateforme de bioinformatique et biostatistique (BIBS)** de l'UMR 7216 Épigénétique et Destin Cellulaire.

depuis 2017 **Coordinateur du comité de pilotage des projets informatiques de l'Institut Jacques Monod**
Suivi des projets. Construction de la stratégie informatique de l'Institut.

depuis 2017 **Correspondant informatique de l'équipe « Mitochondries, métaux et stress oxydatif » à l'Institut Jacques Monod**
Coordination avec le service informatique de l'Institut. Gestion du parc informatique de l'équipe.

depuis 2016 **Responsable du service informatique de l'UFR Sciences du Vivant**
En co-responsabilité avec un autre collègue.
Encadrement de deux techniciens. Gestion d'un parc de **150 machines, 5 serveurs et 7 salles informatiques**. Gestion des stocks. Support aux utilisateurs.

2008 → 2012 **Administrateur système du laboratoire**
En co-responsabilité avec un autre collègue.
Gestion d'un parc d'une trentaine de machines et de 5 serveurs. Commande, installation, configuration et maintenance des machines des utilisateurs. Mise en place d'un wiki pour la documentation de l'équipe. Correspondant informatique de l'unité auprès de l'Inserm.
De 2010 à 2012, **encadrement de deux ingénieurs d'étude** pour l'administration système.

1.4 Animation et communication scientifique

2018 et 2019 Participation aux « Dialogues Entre Chercheurs et Lycéens pour les Intéresser à la Construction des Savoirs » (Declics).

2018 Participation à la Table Ouverte en Bioinformatique (TOBi) à Paris.
Invitation par l'association JeBiF. Retour d'expérience sur le métier de bioinformaticien.

2018 Organisation d'un atelier *Software Carpentry* (2 jours, 20 personnes). Découverte de Bash, Python et git).
Avec Victoria Dominguez Del Angel de l'Institut Français de Bioinformatique.

2011 Organisation du XVIIe congrès du Groupe de Graphisme et Modélisation Moléculaire (GGMM), (3 jours, 120 personnes, 27 conférences, 62 posters).
Avec Sophie Sacquin-Mora, Marc Baaden et Florent Barbault.

2010 Participation à l'opération « 1000 chercheurs parlent d'avenir » créée par le photographe Pierre Maraval.

depuis 2010 Présence sur les réseaux sociaux. @pierrepo sur Twitter

1.5 Financements de projets

2021 *D'inversée à renversée, la classe dans tous ses états comme autant de preuves de concepts des pédagogies actives*
Projet dans le cadre de l'appel à projet « Innovations pédagogiques - Hybridation des formations & Pédagogies innovantes » financé par l'IdEx Université de Paris.
Montant : 2 050 €

2019 *PLASMA : Plateforme d'e-Learning pour l'Analyse de données Scientifiques Massives*
Projet dans le cadre de l'appel à projet « Formation d'excellence / Projets nouveaux cursus et pédagogie innovante » financé par l'IdEx Université de Paris et porté avec Claire Vandiedonck et Sandrine Caburet
Montant : 61 858 €

2019 *PLASMA : Plateforme d'e-Learning pour l'Analyse de données Scientifiques Massives*
Projet financé par l'Ecole Universitaire de Recherche (EUR) GENE et porté avec Claire Vandiedonck et Sandrine Caburet
Montant : 10 000 €

2018 *PLASMA : Plateforme d'e-Learning pour l'Analyse de données Scientifiques Massives*
Projet dans le cadre de l'appel à projet « Trophées franciliens de l'innovation numérique dans le supérieur » (EdTech2018) financé par la Région Ile-de-France et porté avec Claire Vandiedonck et Sandrine Caburet
Montant : 75 000 €

2018 *Live coding pour l'enseignement de la bioinformatique*
Projet dans le cadre du prix « Innovation pédagogique numérique » (IN2018) financé par l'Université Paris Diderot et porté avec Patrick Fuchs
Montant : 2 000 €

2017 *Formation continue « Création, analyse et valorisation de données biologiques omiques »*
Projet dans le cadre de l'appel à projet « Création de nouvelles formations continues » financé par l'Université Paris Diderot et porté avec Gaëlle Lelandais et Bertrand Cosson
Montant : 32 000 €

2015 *Technologies numériques contre le paludisme*
Projet financé par Total E&P Congo
Montant : 20 606 €

2014 *Analyse de données cliniques*
Projet financé par Total E&P Congo
Montant : 22 712 €

1.6 Encadrement d'étudiants

depuis 01/2021 : William Amory, L3 Biologie Informatique, Université de Paris
La loi de Benford en biologie : existence et pertinence

depuis 10/2020 : Thibault Poinsignon, thèse (co-direction avec Gaëlle Lelandais), Université Paris-Saclay
Minomics : modélisation des réseaux biologiques, analyse de données multi-omiques, visualisation à large échelle

01/2020 → 07/2020 : Akram Hecini, M2 Biologie Informatique, Université de Paris
Characterization of the missing peptidome

03/2019 → 07/2019 : Lilian Yang-Crosson, M1 Biologie Informatique, Université Paris Diderot
Analyse de la quantification de peptides par marquage d'isotope léger

01/2017 → 07/2017 : Athénaïs Vaginay, M2 Biologie Informatique, Université Paris Diderot
Banque de données de clichés annotés de gouttes épaisses pour le diagnostic du paludisme et diagnostic automatisé

11/2014 → 07/2016 : Laure Stella Ghoma Linguissi, thèse (co-encadrement avec Francine Ntoumi et Jacques Simpore), Université de Ouagadougou, Burkina Faso
Diagnostic de la tuberculose en République du Congo et *Diversité génétique du VIH et prévention de la transmission du VIH de la mère à l'enfant*

02/2012 → 08/2012 : Marie Ober, M2 Bioinformatique et Biostatistiques, Université Paris Sud (Orsay)
Classification et génération de modèles à haut débit des intégrines

06/2011 → 07/2011 : Clotilde Guyon, M1 Biochimie, Cellules, Cibles Thérapeutiques, Université Paris Diderot
Modélisation des polymorphismes des intégrines $\alpha 2b\beta 3$ dans la thrombasténie de Glanzmann

03/2011 → 04/2011 : Franck Da Silva & Adrien Villain, L3 Biologie Informatique, Université Paris Diderot
Développement d'un programme d'analyse de séquences : XPybaloo
Co-encadrement avec Olivier Bertrand

03/2009 → 05/2009 : Christel Goudot, M1 Biologie Informatique, Université Paris Diderot
Création d'une base de données CMF BASE pour l'analyse de données issues de la cytométrie en flux à haut débit dans l'étude de maladies vaso-occlusives
Co-encadrement avec Julien Picot et Gaëlle Lelandais

01/2009 → 06/2009 : Florence Dol, M2 Biologie Informatique, Université Paris Diderot
Relation séquence / structure dans la famille des sHSP
Co-encadrement avec Delphine Flatters

03/2008 → 06/2008 : Matthieu Almeida, M1 Biologie Informatique, Université Paris Diderot
Création d'une base de données sHSPprotseqDB pour l'analyse des petites protéines de choc thermique
Co-encadrement avec Delphine Flatters

1.7 Résumé des activités d'enseignement

Enseignements en gestion, modélisation et analyse de données biologiques, en protéomique et en informatique (Unix, *shell* Bash, programmation Python, développement *open source*) :

- L1 Sciences du Vivant (250 apprenants), UE « Modélisation mathématique en biologie » †
- L1 Sciences du Vivant (15 apprenants), UE « Introduction à la bioinformatique » (option)
- L3 Sciences du Vivant (170 apprenants), UE « Les omiques »
- M1 Biologie Informatique (30 apprenants), UE « Programmation Python 2 » †
- M1 Magistère Européen de Génétique (45 apprenants), UE « Approches génétiques et génomiques à l'ère des données massives »
- M2 Biologie Informatique (30 apprenants), UE « Programmation 3 et projet tuteuré »
- M1 AIRE Digital Science (18 apprenants), UE « Open source » †*
- formation continue diplôme universitaire « Création, analyse et valorisation de données biologiques omiques » (DU Omiques)‡ (14 apprenants), UE « Gestion de données et reproductibilité des analyses » † et « Automatisation du processus d'analyse de données » †
- formation continue diplôme universitaire en « Bioinformatique intégrative » (DUBii)‡ en partenariat avec l'Institut Français de Bioinformatique (19 apprenants), UE « Environnement Unix » † et « Programmation Python » †

† indique les responsabilités ou co-responsabilités d'UE
‡ indique les responsabilités de formation
* indique les cursus internationaux.

Depuis 2020, je suis membre du groupe de travail *e-learning* de l'Institut Français de Bioinformatique animé par Hélène Chiapello. Nous étudions les dispositifs de formation en ligne pour le développement de compétences de base en Unix, Python et R. Nous avons conçu un premier parcours d'initiation à Unix sur la plateforme Katacoda :
https://www.katacoda.com/ifb-elixirfr/courses/ifb-unix

1.7.1 Actions clés axées sur la pédagogie

2021 — Lauréat de l'appel à projets IdEx « Innovations pédagogiques - Hybridation des formation & Pédagogies innovantes ».
Projet : « D'inversée à renversée, la classe dans tous ses états comme autant de preuves de concepts des pédagogies actives » (2 k€).

2020 — Communication orale aux journées d'étude de l'AIPU : « CertifiENS : le rôle du "passeur accompagnateur" dans la transformation pédagogique ».

2020 Publication d'un article en *preprint* sur les classes inversée « Ten Simple Rules for Implementing a Flipped Classroom ».
https://www.preprints.org/manuscript/202007.0030/v2

2020 Communication orale lors du congrès sur les classes inversées et les pédagogies actives (CL !C 2020) : « Expérimentation de la classe inversée pour l'enseignement de la programmation Python à l'université ».

2019 Certificat de pédagogie CertifiENS, délivré par SAPIENS.
Formation de 2 ans en pédagogie. Participation à 8 ateliers, un accompagnement individuel, une séance d'observation et 2 journées réflexives. Rédaction d'un bilan réflexif.

2018 Prix de l'innovation pédagogique numérique (IN2018) de l'Université Paris Diderot : « Live coding pour l'enseignement de la bioinformatique » (2 k€).

2018 Trophées franciliens de l'innovation numérique dans le supérieur (EdTech2018) : « PLASMA : Plateforme d'e-Learning pour l'Analyse de données Scientifiques Massives » (75 k€).

2017 Création de la formation continue : Diplôme Universitaire « Création, analyse et valorisation de données biologiques omiques ». Appel à projet « création de nouvelles formations continues » (DU Omiques) : 32 k€.
Formation créée et animée avec Bertrand Cosson (Université de Paris) et Gaëlle Lelandais (Université Paris-Saclay).

1.7.2 Focus sur quelques projets pédagogiques

Dans cette partie, je détaille certains projets pédagogiques auxquels j'ai participé.

2018 : Plateforme d'e-Learning pour l'Analyse de données Scientifiques Massives (Plasma)

Plasma (https://plasmabio.org/) est une plateforme d'e-learning dédiée à l'analyse de données. Elle repose sur l'écosystème *open-source* Jupyter avec notamment le *hub* (portail de connexion), le *lab* (interface web d'analyse) et les *notebooks*. Les *notebooks* Jupyter sont des cahiers numériques d'analyse, qui peuvent contenir du texte, des images, des équations mathématiques ou du code informatique dans un ou plusieurs langages de programmation. Ces *notebooks* sont non seulement très pratiques pour analyser des données, mais sont aussi particulièrement conviviaux pour enseigner la programmation ou l'analyse de données via une approche guidée et progressive. Enfin, ces *notebooks* contribuent à améliorer la reproductibilité des analyses.

Ce projet, co-porté avec Claire Vandiedonck et Sandrine Caburet, a été lauréat en 2018 des Trophées franciliens de l'innovation numérique dans le supérieur (EdTech 2018) et a bénéficié du soutien de la région Ile-de-France à hauteur de 75 k€.

Ce projet a également été soutenu par l'Idex de l'Université de Paris, l'EUR GENE et le diplôme universitaire « DU Omiques ». Tous les développements ont été réalisés en partenariat avec la société QuantStack et sont *open-sources* (https://github.com/plasmabio/plasma).

Depuis septembre 2020, deux serveurs d'analyses sont en production et utilisés pour les enseignements du magistère européen de génétique.

Ce projet a également financé le développement d'une extension Jupyter pour le moteur de construction et de visualisation de réseaux Cytoscape. Cette extension, `ipycytoscape`, a été particulièrement appréciée par la communauté et se trouve désormais officiellement supportée par le consortium Cytoscape (https://github.com/cytoscape/ipycytoscape).

2017 : Montage d'une formation continue

Avec Gaelle Lelandais (Université Paris-Saclay) et Bertrand Cosson (Université de Paris), nous avons créé en 2017 le diplôme universitaire « Création, analyse et valorisation de données biologiques omiques » (DU Omiques).

Cette formation continue offre la possibilité d'acquérir une expertise en analyse de données « omiques » avec un objectif opérationnel. Elle a été conçue pour développer des compétences en analyse de données expérimentales haut débit, permettant aux participants de renforcer une équipe scientifique ou un service pour une mission d'analyse de données « omiques ».

Nous avons structuré cette formation en 10 sessions de 2 jours chacune, réparties sur 1 an. Depuis sa création, nous accueillons et formons environ 15 stagiaires chaque année.

Pour le montage de cette formation, nous avons remporté en 2017 un appel à projet de l'Université Paris Diderot pour la création de nouvelles formations d'un montant de 32 k€.

2007 : Ressource libre pour l'apprentissage de la programmation Python

Depuis 2007, je maintiens avec mon collègue Patrick Fuchs un cours en ligne pour l'apprentissage de la programmation Python (https://python.sdv.univ-paris-diderot.fr/). Nous mettons à jour ce cours plusieurs fois par an et l'intégralité du contenu est sous licence libre Creative Commons Attribution - Partage à l'identique (CC BY-SA).

Ce cours reçoit environ 50 000 visites par mois et possède une certaine notoriété dans la communauté francophone.

En 2017, l'éditeur Dunod nous a proposé de publier ce cours sous forme d'un ouvrage dans la collection Sciences Sup. Le livre est paru en 2019 sous le titre « Programmation Python pour les sciences de la vie » [1].

L'intégralité de nos droits d'auteurs est versée à deux associations : Wikimedia France qui s'occupe notamment de l'encyclopédie libre Wikipédia et NumFocus qui soutient le développement de logiciels libres scientifiques et notamment de l'écosystème scientifique autour de Python.

1. https://www.dunod.com/sciences-techniques/programmation-en-python-pour-sciences-vie

Partie 2

Titres et travaux

Les articles en *open access* sont indiqués par le logo

2.1 Publications scientifiques avec comité de lecture

26. Sénécaut N, Alves G, Weisser H, Lignières L, Terrier S, Yang-Crosson L, **Poulain P**, Lelandais G, Yu YK, Camadro JM,
Novel Insights into Quantitative Proteomics from an Innovative Bottom-Up Simple Light Isotope Metabolic (bSLIM) Labeling Data Processing Strategy
Journal of Proteome Research, 20(3) : 1476–1487, (2021).

25. Camadro JM, **Poulain P**,
AutoClassWrapper : a Python wrapper for AutoClass C classification
The Journal of Open Source Software, 4(39) : 1390 (2019).

24. Denecker T*, Durand W*, Maupetit J*, Hébert C, Camadro JM, **Poulain P**†, Lelandais G†,
Pixel : a content management platform for quantitative omics data
PeerJ, 7 :e6623 (2019).

* Ces auteurs ont contribué équitablement à ce travail.
† Ces auteurs ont contribué équitablement à ce travail.

23. Abdollahi N, Albani A, Anthony E, Baud A, Cardon M, Clerc R, Czernecki D, Conte R, David L, Delaune A, Djerroud S, Fourgoux P, Guiglielmoni N, Laurentie J, Lehmann N, Lochard C, Montagne R, Myrodia V, Opuu V, Parey E, Polit L, Privé S, Quignot C, Ruiz-Cuevas M, Sissoko M, Sompairac N, Vallerix A, Verrecchia V, Delarue M, Guérois R, Ponty Y, Sacquin-Mora S, Carbone A, Froidevaux C, Le Crom S, Lespinet O, Weigt M, Abboud S, Bernardes J, Bouvier G, Dequeker C, Ferré A, Fuchs P, Lelandais G, **Poulain P**, Richard H, Schweke S, Laine E, Lopes A,
Meet-U : Educating through research immersion
PLOS Computational Biology, 14(3) : e1005992 (2018).

22. Barnoud J*, Santuz H*, Craveur P, Joseph AP, Jallu V, de Brevern AG†, **Poulain P**†,
PBxplore : a tool to analyze local protein structure and deformability with Protein Blocks
PeerJ, 5 : e4013 (2017).

* Ces auteurs ont contribué équitablement à ce travail.
† Ces auteurs ont contribué équitablement à ce travail.

21. Etoka-Beka MK, Ntoumi F, Kombo M, Deibert J, **Poulain P**, Vouvoungui C, Kobawila SC, Koukouikila-Koussounda F,
Plasmodium falciparum infection in febrile Congolese children : prevalence of clinical malaria 10 years after introduction of artemisinin-combination therapies
Tropical Medicine & International Health, 21 : 1496 (2016).

20. Ghoma Linguissi LS, Ndembi N, Nkenfou CN, **Poulain P***, Ntoumi F*,
HIV-1 Genetic Diversity in the Republic of Congo : Seventeen Years in Review
JSM Microbiology, 3 : 1025 (2015).

* Ces auteurs ont contribué équitablement à ce travail.

19. Ghoma Linguissi LS, Vouvoungui JC, **Poulain P**, Bango Essassa G, Kwedi S, Ntoumi F,
Diagnosis of smear-negative pulmonary tuberculosis based on clinical signs in the Republic of Congo
BMC Research Notes, 8 : 804 (2015).

18. Ghoma Linguissi LS, Ebourombi DF, Sidibe A, Kivouele TS, Vouvoungui JC, **Poulain P**, Ntoumi F,
Factors influencing acceptability of voluntary HIV testing among pregnant women in Gamboma, Republic of Congo
BMC Research Notes, 8 : 652 (2015).

17. Ghoma Linguissi LS, Bisseye C, **Poulain P**, Ntoumi F, Simpore J,
Prevention of Mother-to-Child HIV Transmission (PMTCT) in the Republic of Congo : Challenges to Implementation
Journal of AIDS & Clinical Research, 6 : 503 (2015).

16. Craveur P, Joseph AP, Esque J, Narwani TJ, Noël F, Shinada N, Goguet M, Leonard S, **Poulain P**, Bertrand O, Faure G, Rebehmed J, Ghozlane A, Swapna LS, Bhaskara RM, Barnoud J, Téletchéa S, Jallu V, Cerny J, Schneider B, Etchebest C, Srinivasan N, Gelly J-C, de Brevern AG,
Protein flexibility in the light of structural alphabets
Frontiers in Molecular Biosciences, 2 : 20 (2015).

15. Boyer B, Ezelin J, **Poulain P**, Saladin A, Zacharias M, Robert CH, Prévost C,
An Integrative Approach to the Study of Filamentous Oligomeric Assemblies, with Application to RecA
PLoS ONE, 10 : e0116414 (2015).

14. Jallu V*, **Poulain P***, Fuchs PFJ, Kaplan C, de Brevern AG,
Modeling and molecular dynamics simulations of the V33 variant of the integrin subunit β3 : Structural comparison with the L33 (HPA-1a) and P33 (HPA-1b) variants
Biochimie, 105 : 84-80 (2014).

* Ces auteurs ont contribué équitablement à ce travail.

13. Craveur P, Joseph AP, **Poulain P**, de Brevern AG, Rebehmed J,
Cis–trans isomerization of omega dihedrals in proteins
Amino Acids, 45 : 279-289 (2013).

12. Jallu V, Bertrand G, Bianchi F, Chenet C, **Poulain P**, Kaplan C,
The αIIb p.Leu841Met (Cab3a+) Polymorphism Results in a New Human Platelet Alloantigen Involved in Neonatal Alloimmune Thrombocytopenia
Transfusion, 53 : 554-563 (2013).

11. Jallu V*, **Poulain P***, Fuchs PFJ, Kaplan C, de Brevern AG,
Modeling and Molecular Dynamics of HPA-1a and -1b Polymorphisms : Effects on the Structure of the β3 Subunit of the αIIbβ3 Integrin
PLoS ONE, 7 : e47304 (2012).

* Ces auteurs ont contribué équitablement à ce travail.

10. Dall'Olio GM, Marino J, Schubert M, Keys KL, Stefan MI, Gillespie CS, **Poulain P**, Shameer K, Sugar R, Invergo BM, Jensen LJ, Bertranpetit J, Laayouni H,
Ten Simple Rules for Getting Help from Online Scientific Communities
PLoS Computational Biology, 7 : e1002202 (2011).

9. Saladin A, Amourda C, **Poulain P**, Férey N, Baaden M, Zacharias M, Delalande O, Prévost C,
Modeling the early stage of DNA sequence recognition within RecA nucleoprotein filaments
Nucleic Acids Research, 38 : 6313-6323 (2010).

8. **Poulain P**, Gelly J-C, Flatters D,
Detection and Architecture of Small Heat Shock Protein Monomers
PLoS ONE 5 : e9990 (2010).

7. Saladin A, Fiorucci S, **Poulain P**, Prévost C, Zacharias M,
PTools : an opensource molecular docking library
BMC Structural Biology 9 : 27 (2009).

6. Calvo F, **Poulain P**,
Transitions between secondary structures in isolated polyalanines
The European Physical Journal D 51 : 15-23 (2009).

5. **Poulain P**, Saladin A, Hartmann B, Prévost C,
Insights on protein-DNA recognition by coarse grain modelling
Journal of Computational Chemistry 29 : 2582-2592 (2008).

4. **Poulain P**, Calvo F, Antoine R, Broyer M, Dugourd P,
Competition between secondary structures in gas phase polyalanines
Europhysics Letters 79 :66003 (2007).

3. **Poulain P**, Calvo F, Dugourd P, Antoine R, Broyer M,
Performances of Wang-Landau algorithms for continuous systems
Physical Review E 73 : 056704 (2006).

2. Antoine R, Broyer M, Chamot-Rooke J, Dedonder C, Desfrancois C, Dugourd P, Gregoire G, Jouvet C, Onidas D, **Poulain P**, Tabarin T, van der Rest G,
Comparison of the fragmentation pattern induced by collisions, laser excitation and electron capture. Influence of the initial excitation
Rapid Communications in Mass Spectrometry 20 : 1-5 (2006).

1. **Poulain P**, Antoine R, Broyer M, Dugourd P,
Monte Carlo simulations of flexible molecules in a static electric field : electric dipole and conformation
Chemical Physics Letters 400 : 1-6 (2005).

2.2 Publications scientifiques sans comité de lecture

Jallu V, **Poulain P**, Kaplan C, de Brevern AG,
3D protein structure modeling : A tool to provide insight into the platelet alloimmune response
Transfusion Today 86 : 10-11 (2011).

2.3 Livre

Fuchs P, Poulain P,
Programmation Python pour les sciences de la vie,
Dunod, Paris, (2019). EAN 9782100796021

2.4 Communications orales

2.4.1 Communications orales dans des conférences nationales

2020/10/30-2020/11/01 - Paris - France (*virtuel*)
Congrès des classes inversées et des pédagogies actives (CLIC) 2020
Poulain P
Expérimentation de la classe inversée pour l'enseignement de la programmation Python [vidéo]

2018/11/27 - Paris - France
Journée « Interopérabilité et pérennisation des données de la recherche : comment "FAIR" en pratique ? Retours d'expériences »
Poulain P
Retour d'expériences sur la publication de données en biologie [vidéo]

2017/12/19 - Rennes - France
Journée Python et Data Science
Poulain P
Lutter contre le paludisme avec Python

2017/06/12-13 - Paris - France
PyParis 2017
Poulain P
Using Python to fight Malaria

2017/06/12-13 - Evry - France
Atelier « Modélisation Gros-Grains pour la Biologie et la Matière Molle »

Poulain P
Simulations d'assemblage protéine/ADN

2007/01/31-2007/02/02 - Montpellier - France
journée Jeune Chercheur Calculant au CINES
Poulain P, Calvo C, Broyer M, Antoine R, Dugourd P
Simulation de protéines dans les ensembles généralisés

2006/09/19 - Montpellier - France
journée Jeune Chercheur Calculant au CINES
Poulain P, Calvo C, Broyer M, Antoine R, Dugourd P
Simulation de protéines dans les ensembles généralisés

2.4.2 Séminaires de laboratoire

2007/07/10 - Breme - Allemagne
Computational Biology Group, Jacobs University
Poulain P
Docking simulations of flexible proteine–DNA complexes

2007/05/31 - Paris - France
Laboratoire de Physique Théorique de la Matière Condensée (LPTMC)
Poulain P
Simulations de peptides en phase gazeuse par la méthode Wang-Landau

2006/04/06 - Paris - France
Laboratoire de Biochimie Théorique
Poulain P, Calvo C, Broyer M, Antoine R, Dugourd P
Conformations de biomolécules : simulations dans les ensembles généralisés

2.5 Communications par affiche

L'auteur qui a présenté le poster est indiqué en gras.

2020/10/12-16 - Berlin - Allemagne (*virtuel*)
JupyterCon 2020
Tuloup J, Vandiedonck C, Caburet S, **Poulain P**
Plasma : versatile e-learning platform powered by The Littlest JupyterHub

2020/06/30-2020/07/03 - Montpellier - France (*virtuel*)
Journées Ouvertes de Biologie, Informatique et Mathématique (JOBIM) 2020
Tuloup J, Vandiedonck C, **Poulain P**, Caburet S
Plasma : e-learning platform for massive data analysis

2020/06/30-2020/07/03 - Montpellier - France (*virtuel*)
Journées Ouvertes de Biologie, Informatique et Mathématique (JOBIM) 2020
Denecker T, Sénécaut N, Poulain P, Lelandais G, Camadro JMS
Systematic Analysis of Protein Post-translational Modifications at a Proteomic Scale in the pathogenic yeast Candida albicans

2018/07/03-06 - Marseille - France
Journées Ouvertes de Biologie, Informatique et Mathématique (JOBIM) 2018
Denecker T, Durand W, Maupetit J, Hébert C, Camadro JM, Poulain P, Lelandais G
Pixel : une solution Open Source pour l'annotation, le stockage, l'exploration et l'intégration des résultats d'analyses de données multi-omiques en biologie
Prix SFBI du meilleur poster.

2015/11/19-21 - Munich - Allemagne
Joint Annual Meeting German Society of Infectious Diseases (DGI) and German Center for Infection Research (DZIF)
Etoka-Beka MK, Kombo M, Deibert J, Poulain P, Vouvoungui C, Koukouikila-Koussounda F, Ntoumi F
Plasmodium falciparum infection in febrile Congolese children : prevalence of clinical malaria and influence of sickle cell trait

2013/06/11-13 - Paris - France
XXVIème congrès de la Société Française de Transfusion Sanguine
Jallu V, Bertrand G, Bianchi F, Chenet C, Poulain P, Kaplan C
Un nouvel alloantigène plaquettaire (HPA-27bw) impliqué dans l'alloimmunisation materno-foetale : le variant Met841 de la sous-unité αIIb.

2011/06/28 - Paris - France
Journées Ouvertes de Biologie, Informatique et Mathématique (JOBIM) 2011
Poulain P, Jallu V, Kaplan C, de Brevern AG
In Silico Insights into the Platelet Alloimmune Response to $\alpha IIb\beta 3$ Polymorphisms

2011/05/30 - La Rochelle - France
XVIIe congrès du Groupe de Graphisme et Modélisation Moléculaire (GGMM)
Saladin A, Amourda C, Poulain P, Férey N, Baaden M, Zacharias M, Delalande O, **Prévost C**
Modélisation de l'intermédiaire initial de reconnaissance de séquence dans les nucléofilaments de recombinaison homologue

2011/05/30 - La Rochelle - France
XVIIe congrès du Groupe de Graphisme et Modélisation Moléculaire (GGMM)
Boyer B, Poulain P, Zacharias Z, Prévost C
L'amarrage flexible protéine-ADN en modèle gros grains

2011/05/30 - La Rochelle - France
XVIIe congrès du Groupe de Graphisme et Modélisation Moléculaire (GGMM)

Fiorucci S, Saladin A, Poulain P, Prévost C, Zacharias M, Nafati N, Golebiowski J, Antonczak S
Nouvelles approches de docking pour la compréhension des mécanismes de l'olfaction

2010/09/26 - Ghent - Belgium
ECCB10, the 9th European Conference on Computational Biology
Poulain P, Saladin A, Fiorucci S, Zacharias M, Prévost C
PTools : an open source molecular docking library

2010/09/26 - Ghent - Belgium
ECCB10, the 9th European Conference on Computational Biology
Poulain P, Gelly JC, **Flatters D**
Detection and architecture of small heat shock protein monomers

2010/04/06 - Paris - France
Rencontres de la Montagne Sainte Geneviève (fondation Pierre-Gilles de Gennes)
Saladin A, Amourda C, Poulain P, Férey N, Baaden M, Zacharias M, Delalande O, **Prévost C**
Sequence Recognition within RecA Nucleoprotein Filaments : A Molecular Modelling Study

2009/06/09 - Nantes - France
Journées Ouvertes de Biologie, Informatique et Mathématique (JOBIM) 2009
Almeida M , Poulain P, Etchebest C, **Flatters D**
sHSPprotseqDB : a database for the analysis of small Heat Shock Proteins

2009/05/05 - Mittelwihr - France
XVIe congrès du Groupe de Graphisme et Modélisation Moléculaire (GGMM)
Almeida M, **Poulain P**, Etchebest C, Flatters D
Base de données pour l'analyse des petites protéines de choc thermique (sHSP)

2009/03/18 - Sophia Antipolis - France
Flexibilité et Reconnaissance Biologique : de la biophysique aux modèles de données
Saladin A, Poulain P, Zacharias M, **Prévost C**
Modelling Sequence Recognition in Homologous Recombination

2007/06/07 - Paris - France
Journées de simulation numérique (JSNUM)
Poulain P, Saladin A, Prévost C
Simulations d'assemblages flexibles protéine/ADN

2007/05/02 - Autrans - France
XVe congrès du Groupe de Graphisme et Modélisation Moléculaire (GGMM)
Poulain P, Saladin A, Prévost C
Simulations d'assemblages flexibles protéine/ADN

2006/11/28–2006/12/01 - Orsay - France
Atelier « Sampling paths in molecular simulation : algorithms for phase transition, reactivity

and kinetics »

Poulain P, Calvo F, Broyer M, Antoine R, Dugourd P

Generalized ensembles simulations of gas-phase polyalanines

Partie 3

Copie du diplôme de doctorat

RÉPUBLIQUE FRANÇAISE

MINISTÈRE DE L'ÉDUCATION NATIONALE, DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE

UNIVERSITÉ LYON 1

DIPLÔME DE DOCTEUR

Vu le code de l'éducation, et notamment son article L.613-1

Vu le décret n° 84-573 du 5 juillet 1984 modifié relatif aux diplômes nationaux de l'enseignement supérieur

Vu l'arrêté du 25 avril 2002 relatif aux études doctorales

Vu le procès-verbal du jury attestant que l'intéressé a soutenu, le 3 juillet 2006 une thèse portant sur le sujet suivant :
"Structure et dynamique de protéines isolées : approches statistiques",
devant un jury présidé par JEAN-LOUIS BARRAT, Professeur des Universités et composé de FLORENT CALVO, Chargé de Recherche CNRS, PHILIPPE DUGOURD, Chargé de Recherche CNRS, CHRISTOPHE JOUVET, Directeur de Recherche, RICHARD LAVERY, Directeur de Recherche CNRS, YVES HENRI SANEJOUAND, Chargé de Recherche CNRS

Vu la décision dudit jury prononçant l'admission de l'intéressé

le Diplôme de docteur en PHYSIQUE

est conféré à **M. PIERRE POULAIN**

né le 21 janvier 1979 à ST- CLOUD (092) pour en jouir avec les droits et prérogatives qui y sont attachés.

Fait à Lyon, le 1 mars 2007

Le titulaire

Le Président

Lionel COLLET

Le Recteur d'Académie,
Chancelier des universités

Alain MORVAN

Académie de Lyon
Rectorat

N° LYONI 6153882
/2007200506970

Partie 4

Synthèse des travaux de recherche

4.1 Résumé de mon parcours scientifique

Un début de carrière en France (–2012)

Après un diplôme d'ingénieur chimiste obtenu en 2003 à l'École Nationale Supérieure de Chimie de Montpellier et un *Master of Science in Cheminformatics* à UMIST (Manchester, Royaume-Uni) la même année, j'ai poursuivi mes études scientifiques et mon initiation à la recherche par une thèse de physique à l'Université Claude Bernard Lyon I sous la direction de Philippe Dugourd. Pendant ces 3 ans au sein de l'équipe « Dipôle Électrique, Biomolécules et Agrégats » du Laboratoire de spectrométrie ionique et moléculaire (Lasim), j'ai exploré la conformation de peptides en phase gazeuse par des approches de physique statistique et des simulations numériques, notamment Monte Carlo.

En 2006, j'ai rejoint Chantal Prévost au Laboratoire de Biochimie Théorique (LBT) situé à l'Institut de Biologie Physico-Chimique (IBPC), à Paris, pour un post-doc CNRS d'un an. Je me suis attaché cette fois à construire un modèle gros d'ADN pour des simulations d'amarrage moléculaire protéine-ADN.

En 2007, je suis recruté comme maître de conférences à l'Université Paris Diderot. J'ai été accueilli dans le laboratoire « Équipe de Bioinformatique Génomique et Moléculaire » (EBGM) dirigé par Catherine Etchebest. En 2008, une partie du laboratoire a été transférée à l'Institut National de la Transfusion Sanguine (INTS), dans l'unité Inserm UMR-S 665 dirigé par Yves Colin. Au sein de l'équipe Dynamique des Structures et des Interactions des Macromolécules Biologiques (DSIMB), j'ai poursuivi des activités de recherche en bioinformatique structurale. J'ai notamment étudié des systèmes biologiques impliqués dans des pathologies (petites protéines de choc thermique, intégrines α2bβ3) par des approches d'analyse de séquences ou de simulations de dynamique moléculaire.

Puis quatre ans au Congo (2012–2016)

En 2012, je me suis mis en disponibilité pour suivre mon épouse mutée à Pointe-Noire, en République du Congo. Ce changement de vie radical m'a donné l'occasion de travailler dans l'industrie pétrolière, d'abord comme responsable de projets informatiques puis comme analyste de données pour le système d'information géographique de la société Total Exploration & Production Congo.

J'ai également maintenu une collaboration active avec l'équipe DSIMB en France et poursuivi certains projets de recherche.

Malgré un tissu scientifique congolais peu lisible, j'ai réussi à prendre contact avec un laboratoire de recherche, situé à Brazzaville, à 500 km de Pointe-Noire. En 2014, j'ai intégré la Fondation Congolaise pour la Recherche Médicale (FCRM) dirigée par Francine Ntoumi. Dans un premier temps, j'ai encadré l'activité de recherche d'une étudiante en thèse ce qui nous a conduit à publier quatre articles. J'ai également collaboré avec le biostatisticien de la fondation sur l'organisation et l'analyse des données cliniques. Enfin, en 2015, j'ai initié, toujours pour la FCRM, plusieurs projets innovants reposant sur l'utilisation des technologies numériques pour lutter contre le paludisme. J'ai ainsi coordonné l'activité de Fongwama, la plateforme congolaise de développement libre, constituée de 6 jeunes développeurs congolais. Nous avons produit deux

applications. La première, `EduPalu`, est une application d'information et d'éducation sur le paludisme au Congo. La seconde, `DensiPara`, est une application de calcul de la densité parasitaire pour le diagnostic du paludisme.

Un retour en France avec de nouvelles idées et une place à trouver (2016–)

Mon expérience de 4 ans en République du Congo a été une révélation, tant par la réflexion sur mes pratiques pédagogiques que la découverte de nouveaux centres d'intérêt scientifiques, notamment autour des données en biologie. En suivant avec beaucoup d'intérêt l'évolution de la bioinformatique et plus généralement de la biologie, j'ai constaté l'ampleur que prennent aujourd'hui les techniques haut-débit dites « omiques ».

L'évolution de la protéomique m'a particulièrement marqué. En effet, les technologies employées actuellement en spectrométrie de masse permettent désormais d'analyser des protéines de plus en plus grosses et à très haut débit. Les données ainsi produites gagnent à la fois en qualité et en quantité. Cette évolution conduit à utiliser aujourd'hui la spectrométrie de masse comme outil de diagnostic pour certaines pathologies.

À mon retour en France, j'ai intégré en janvier 2017, l'équipe « Mitochondries, métaux et stress oxydatif » dirigée par Jean-Michel Camadro à l'Institut Jacques Monod (UMR CNRS 7592). Ce changement de laboratoire avait pour objectif de me confronter aux données « omiques », notamment protéomiques, en bénéficiant de la proximité de la plateforme de protéomique / spectrométrie de masse également dirigée par Jean-Michel Camadro et ainsi de pouvoir bénéficier d'un environnement scientifique riche et stimulant.

Ma première contribution a consisté à développer le cadre théorique pour quantifier des peptides marqués par la technique du *SLIM-labeling* dans le cas d'auxotrophie (détaillée plus loin).

Dans les prochaines rubriques, je vais détailler mes activités de recherche depuis ma thèse (2003).

4.2 Exploration de la structure et de la dynamique de protéines isolées par des approches statistiques

Les protéines sont des molécules omniprésentes chez les êtres vivants. Elles interviennent à tous les stades du fonctionnement d'un organisme. Les fonctions des protéines sont associées à des structures bien spécifiques, adoptées spontanément. Réciproquement, un mauvais repliement peut conduire à une mauvaise fonction, comme c'est le cas, par exemple, pour la maladie neurodégénérative d'Alzheimer, où des protéines solubles sont converties en forme insoluble, conduisant à une mauvaise structure globale et à un comportement pathogène.

Il apparaît donc important de pouvoir comprendre et simuler le repliement des protéines. Cependant, le nombre de conformations possibles pour une protéine est extrêmement grand et rend ce problème très complexe. Les simulations conventionnelles, comme la méthode Monte Carlo ou la dynamique moléculaire dans des conditions classiques (ensemble canonique), ne permettent pas une exploration globale de la surface d'énergie et se retrouvent piégées dans des minima locaux de cette surface. Il faut donc développer des simulations dans des ensembles dits « généralisés » qui permettent une marche aléatoire sur la surface de potentiel avec un franchissement des barrières d'énergie beaucoup plus efficace qu'avec les méthodes conventionnelles.

Mon travail de thèse a consisté à étudier théoriquement les propriétés thermodynamiques de polypeptides en phase gazeuse avec pour objectif une meilleure compréhension des mécanismes fondamentaux impliqués dans le repliement des protéines. Une approche statistique basée sur des algorithmes Monte Carlo dans les ensembles généralisés, comme le Monte Carlo d'échange de répliques ou la méthode Wang-Landau, a été utilisée pour échantillonner le paysage énergétique de ces systèmes complexes. Les peptides étudiés étaient constitués de 2 à 20 acides aminés. Les simulations ont été réalisées en étroite interaction avec les avancées expérimentales du groupe. Nous avons ainsi tenté de comprendre l'influence de la structure secondaire sur les mécanismes de photofragmentation, le rôle de l'entropie dans la stabilisation des feuillets beta à température ambiante et l'effet d'un champ électrique intense sur la conformation de peptides.

D'un point de vue méthodologique, nous avons tout d'abord tenté d'améliorer la méthode Wang-Landau, algorithme d'échantillonnage visant à construire par itérations successives la densité d'états microcanonique en pénalisant les états au fur et à mesure qu'ils sont visités. En particulier, nous avons proposé une modification de cette méthode de façon à réaliser une meilleure exploration du paysage énergétique des systèmes à degrés de liberté continus comme les agrégats d'atomes ou les protéines. Il s'agit, d'une part, de faire évoluer le facteur pénalisant, dit facteur de modification, de la même manière que la température dans une simulation en recuit simulé. D'autre part, de construire la densité d'états à deux dimensions, à savoir en énergie et selon une coordonnée de réaction supplémentaire. Enfin, l'algorithme Wang-Landau a montré une rapidité de convergence supérieure par rapport au Monte Carlo d'échange de réplique sur des systèmes complexes. Ce travail a conduit à une publication (#3).

En collaboration avec des expérimentateurs, nous nous sommes également intéressés à la

photofragmentation de peptides en phase gazeuse. Les mécanismes de fragmentation diffèrent suivant le type d'excitation. Cette dernière peut être une excitation globale sur toute la molécule lors de la collision avec des atomes d'hélium ou localisée sur un chromophore dans le cas d'une excitation due à une impulsion laser. Dans ce dernier cas, la dissociation résultante est fortement influencée par la structure adoptée par le peptide à cause du couplage important entre un état excité et un état dissociatif (publication #2).

Enfin, nous avons simulé les propriétés de polyalanines neutres et isolées. Nous avons observé que :

— À basse température, les structures en hélice α sont stabilisées car elles présentaient le minimum d'énergie potentielle.
— À température intermédiaire, les géométries de type feuillet β sont favorisées entropiquement.
— Enfin, à température élevée, des structures désordonnées, plutôt étirées, sont majoritairement obtenues (Fig. 4.1)

Nous avons également montré que ces résultats étaient en accord avec les mesures expérimentales de dipôle électrique effectuées dans l'équipe. Ce travail a donné lieu à deux publications (#4, #6).

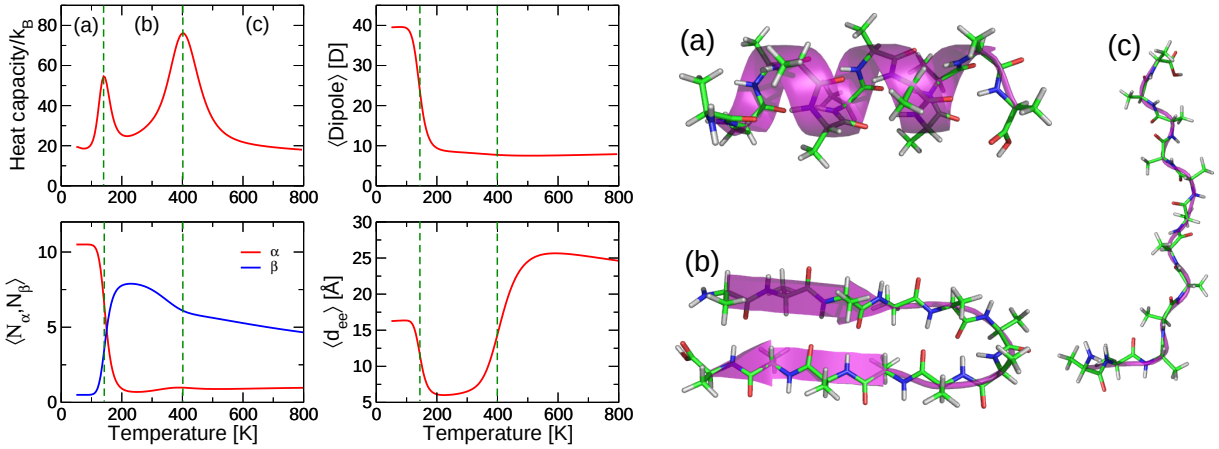


FIGURE 4.1 – *Peptide Ala_{12} en conformation (a) hélice α, (b) feuillet β et (c) structure désordonnée. Ces trois types de structures secondaires ont été mis en évidence par la double transition observée sur la courbe de capacité calorifique. Chaque structure a été caractérisée par les variations avec la température du dipôle électrique, du nombre de résidu en conformation α, $\langle N_\alpha \rangle$, et β, $\langle N_\beta \rangle$, ainsi que les variations de la distance bout-à-bout, $\langle d_{ee} \rangle$.*

Pour terminer, nous avons aussi étudié l'influence théorique d'un champ électrique statique sur les différents éléments de structure secondaire de polyalanines. Le fort dipôle des géométries en hélice α contribuait, sous l'effet du champ électrique, à stabiliser ces conformations au détriment des feuillets β ou des structures désordonnées. Bien que les intensités de champ électrique employées (10^8 V/m) ne puissent pas être mises en œuvre expérimentalement, elles pourraient néanmoins se retrouver localement à proximité de certains ions métalliques dans le milieu biologique (publication #1).

4.3 Modélisation gros grain de l'ADN et amarrage moléculaire

Lors de mon post-doc au Laboratoire de Biochimie Théorique (2006–2007), sous la direction de Chantal Prévost, je me suis intéressé au problème de l'amarrage moléculaire flexible ADN/protéine. La complexation d'une molécule d'ADN avec une protéine intervient dans de nombreux mécanismes biologiques comme le stockage de l'ADN en chromosomes, le contrôle de la transcription ou bien encore la réparation d'ADN endommagé. Être capable de modéliser les interactions existantes dans ces complexes est d'autant plus important que peu de structures sont résolues expérimentalement. L'amarrage d'une protéine à une molécule d'ADN ne se résume pour autant pas à un mécanisme rigide clef/serrure. En effet, des déformations, tant de la protéine que de l'ADN, sont observées dans bon nombre de complexes (Figure 4.2(a)).

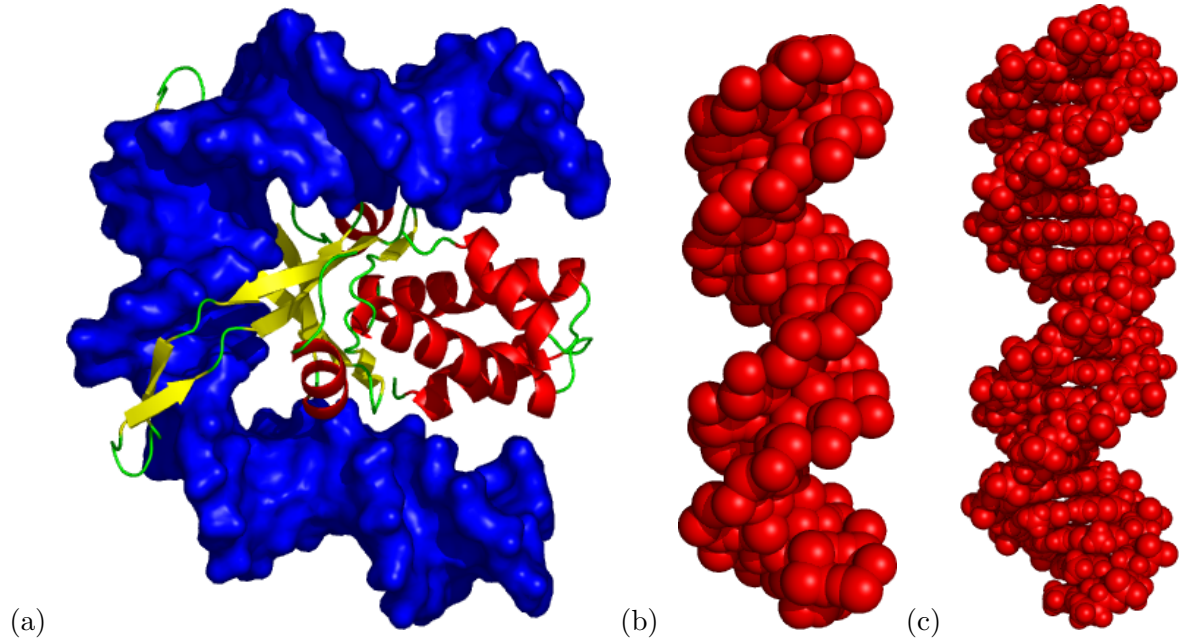


FIGURE 4.2 – *(a) Complexe entre un facteur d'intégration et une molécule d'ADN représentée en bleu par sa surface accessible (code pdb 1IHF). L'ADN est ici fortement déformé. (b) Modèle gros grain de l'ADN B canonique comparé au modèle atomique (c).*

Pour ce projet, j'ai développé un modèle gros grain d'ADN pour des simulations d'amarrage moléculaire protéine / ADN. Le modèle gros grain est un modèle moléculaire où plusieurs atomes sont remplacés par une bille virtuelle, appelée aussi gros grain (Figures 4.2(b) et (c)). L'avantage de ce genre de modèle est de pouvoir accélérer les simulations numériques d'amarrage protéine / ADN. Ce travail a donné lieu à une publication (#5).

Parallèlement à la mise en place de ce modèle gros grain, j'ai participé au développement d'un logiciel pour l'amarrage moléculaire, notamment avec des partenaires gros grain. Cette bibliothèque *open-source* s'appelle `PTools` et est disponible sur GitHub (`https://github.com/ptools/ptools`). Le développement de `PTools` et ses applications ont donné lieu à 3 publications (#7, #9, #15).

4.4 Relation séquence / structure dans la famille des petites protéines de choc thermique

Lors de mon arrivée dans le laboratoire « Équipe de Bioinformatique Génomique et Moléculaire », j'ai collaboré avec Delphine Flatters sur la modélisation des petites protéines de choc thermique (sHSP). Ces protéines, dites quasi-chaperones, interviennent dans l'organisme dans la prévention de l'agrégation irréversible des protéines dénaturées.

La relation séquence / structure est très mal connue chez cette famille de protéines. De nombreuses séquences sont disponibles alors que très peu de structures sont connues. Nous avons étudié un domaine caractéristique de cette famille de protéines, le *Alpha Crystallin Domain* (ACD). En utilisant un jeu de séquences très bien annotées et les structures connues à l'époque (voir Figure 4.3), nous avons construit un profil de l'ACD et ainsi pu proposer une méthode originale de détection et d'annotation de ce domaine sur près de 4500 séquences de sHSP. Nous avons montré que certains résidus étaient extrêmement conservés dans cette famille. Nous avons également identifié une région de l'ACD particulièrement importante pour la formation de la structure quaternaire des sHSP.

Ce travail a donné lieu à une publication (#8).

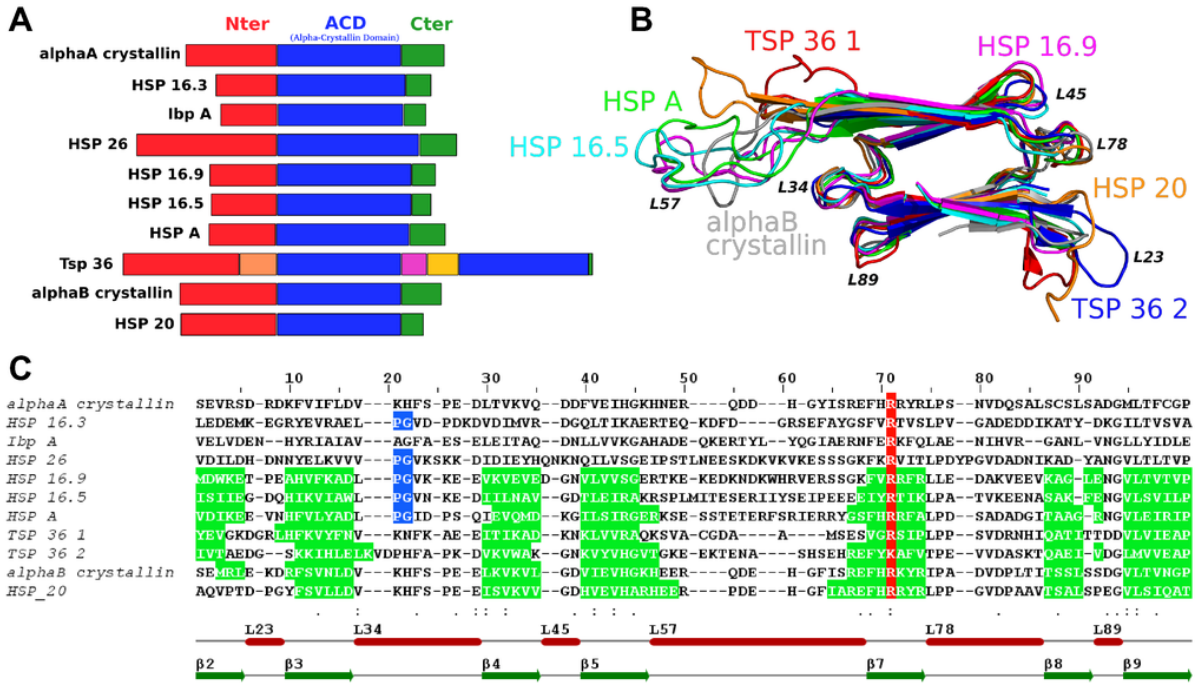


FIGURE 4.3 – Séquences très bien annotées (A) et structures (B) disponibles utilisées pour l'alignement multiple (C) qui a servi à la constructure du profil de l'ACD des sHSP. (Figure tirée de la publication #8)

4.5 Impact des variants du système HPA sur la structure et la dynamique de l'intégrine α2bβ3

Le laboratoire d'immunologie plaquettaire de l'INTS, outre une activité de diagnostic et de référence, développe une activité de recherche sur les causes moléculaires d'un certain nombre de pathologies plaquettaires. Parmi celles-ci, les thrombopénies néonatales alloimmunes (TNAI) résultent de la destruction des plaquettes foetales ou néonatales par des anticorps maternels spécifiques des alloantigènes plaquettaires humains (HPA) hérités du père.

En collaboration avec Vincent Jallu (du laboratoire d'immunologie plaquettaire) et Alexandre de Brevern (de l'équipe DSIMB, laboratoire Inserm U665), nous avons essayé de comprendre l'impact du polymorphisme HPA-1 sur la structure et la dynamique de la sous-unité β3 de l'intégrine α2bβ3.

Nous avons montré qu'une modification localisée de la flexibilité de la sous-unité β3 pouvait avoir des conséquences importantes sur le comportement global de la protéine. En particulier, une augmentation de la flexibilité pourrait expliquer l'augmentation de la capacité d'adhésion des plaquettes portant le polymorphisme HPA-1b et le risque thrombotique associé (voir Figure 4.4).

Ce travail a donné lieu à trois publications (#11, #12, #14).

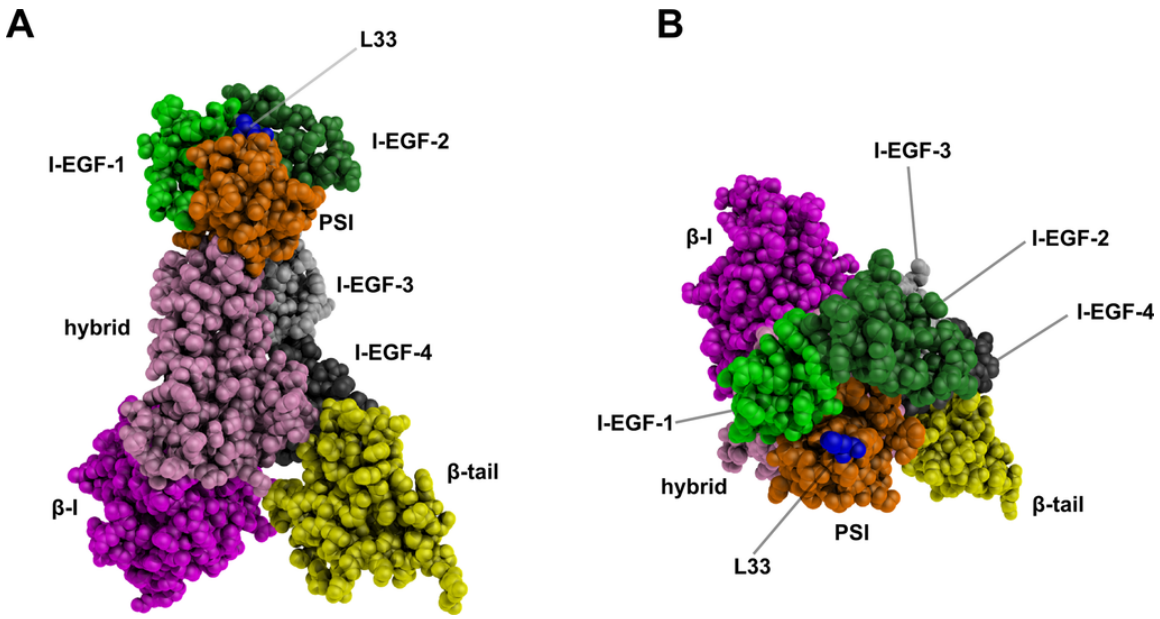


FIGURE 4.4 – Modèle de la sous-unité β3 de l'intégrine α2bβ3. Vue de côté (A) et du dessus (B). Chaque domaine est représenté par une couleur différente. Le résidu L33 (en bleu) correspond au polymorphisme HPA-1 impliqué dans des thrombopénies néonatales alloimmunes. (Figure tirée de la publication #11)

4.6 Modélisation par blocs protéiques

L'étude de la dynamique de protéines m'a conduit à m'intéresser à la conformation locale des résidus et à leur dynamique. L'approche par « blocs protéiques » telle que conceptualisée par Alexandre de Brevern [1] permet de modéliser la conformation tridimensionelle locale du

squelette peptidique sous la forme d'une séquence unidimensionelle de blocs protéiques (*Protein Blocks*, PBs). En principe, n'importe quelle conformation d'acide aminé peut être modélisée par un des 16 PBs disponibles et représentés dans la Figure 4.5.

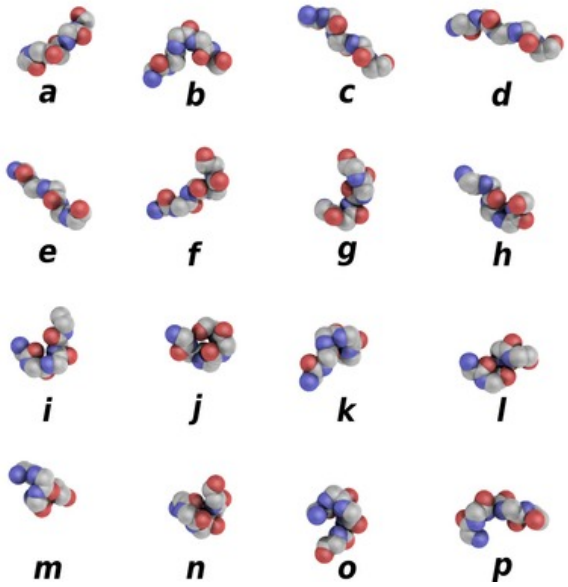


FIGURE 4.5 – Représentation schématique des 16 blocs protéiques, nommés de *a* à *p*.

Pour faciliter l'assignation des blocs protéiques à partir des structures tridimensionnelles des protéines ou de trajectoires de dynamique moléculaire, nous avons développé le logiciel `PBxplore`. `PBxplore` est un logiciel *open-source*, développé en Python, disponible sur PyPI et sur GitHub (https://github.com/pierrepo/PBxplore). Ce travail a donné lieu à la publication #22 en co-dernier auteur.

4.7 Collaboration longue distance

Mon expatriation à Pointe-Noire en République du Congo (2012–2016) n'a pas été simple scientifiquement, car l'organisation de la recherche congolaise n'est pas très lisible et le peu de laboratoires existants souffrent cruellement de moyens.

Mes premières activités professionnelles au Congo se sont donc faites dans l'industrie pétrolière. J'avais néanmoins décidé de dédier bénévolement une partie de mon temps au maintien de collaborations avec mes collègues français. Ce fut un défi, notamment à cause d'une connexion internet régulièrement défaillante. Néanmoins, 4 publications virent le jour de ces collaborations longue distance (#13, #14, #15, #16).

4.8 Traitement et analyse de données cliniques

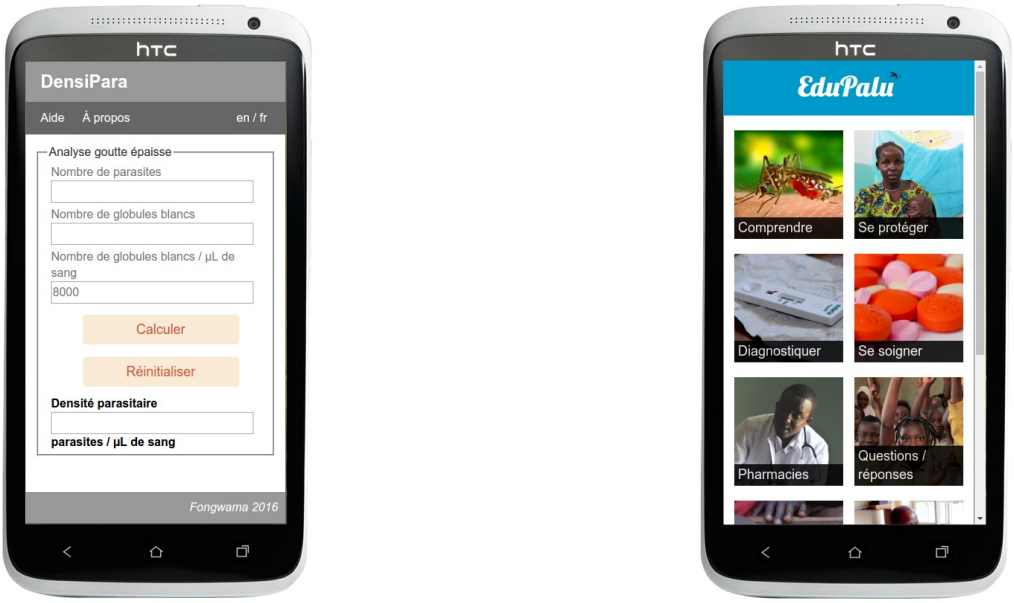
En 2014, j'ai rencontré Francine Ntoumi, directrice générale et fondatrice de la Fondation Congolaise pour la Recherche Médicale (FCRM), située à Brazzaville, à 500 km de Pointe-Noire. Les activités de recherche de la FCRM sont centrées sur les pathologies ayant un fort impact sanitaire au Congo : le paludisme, le VIH/Sida, la tuberculose et les maladies diarrhéiques. Les

approches méthodologiques utilisées sont essentiellement des études de cohortes abordées sous l'angle social, clinique et moléculaire (PCR notamment).

Dans un premier temps, j'ai encadré le travail de Laure Stella Ghoma Linguissi, étudiante en thèse à l'Université de Ouagadougou au Burkina Faso, et qui, ayant terminé son étude sur le VIH au Burkina, est rentrée au Congo pour poursuivre ses activités de recherche sur le VIH et notamment la transmission du VIH de la mère à l'enfant, mais aussi sur la tuberculose, première maladie opportune du VIH.

Mon travail avec Laure a consisté à coordonner l'activité de recherche ainsi que les analyses des données cliniques, en étroite interaction avec Christevie Vouvoungui, biostatisticien à la FCRM. Ce travail a donné lieu à 4 publications (#17, #18, #19, #20) dont une en co-dernier auteur.

4.9 Technologies numériques pour lutter contre le paludisme



(A) **(B)**

FIGURE 4.6 – Aperçu de l'application `DensiPara` (A) et `EduPalu` (B) depuis un téléphone mobile.

En 2015, la FCRM a établi un partenariat avec Fongwama, la plateforme congolaise de développement libre, constituée de 6 jeunes développeurs congolais bénévoles. Une des premières missions de cette plateforme a été la lutte contre le paludisme par les technologies numériques. J'ai supervisé le travail de Fongwama, de juin 2015 à juillet 2016. Nous avons ainsi développé deux applications : `DensiPara` et `EduPalu` (voir copies d'écran Figure 4.6). J'ai assuré la coordination scientifique du projet avec la FCRM et la gestion opérationnelle du projet avec l'entreprise Skepsos.

DensiPara

`DensiPara` est une application de calcul de la densité parasitaire pour le diagnostic du paludisme. Cette application est à destination des techniciens de laboratoire qui diagnostiquent

le paludisme. Elle facilite le calcul de la densité parasitaire qui évalue la quantité de parasites dans le sang. Elle implémente également les recommandations de l'OMS pour les bonnes pratiques de diagnostic.

Cette application, *open-source* et gratuite, est disponible sur le web (https://fongwama.github.io/DensiPara/) et pour les téléphones Android (https://play.google.com/store/apps/details?id=com.fcrm.densipara&hl=fr).

En juin 2016, j'ai coanimé avec Félix Koukouikila-Koussounda, chercheur à la FCRM, un atelier de formation sur `DensiPara` auprès d'une vingtaine de techniciens de laboratoire congolais.

EduPalu

`EduPalu` est une application d'information et d'éducation sur le paludisme au Congo. Elle se destine à l'ensemble de la population.

Cette application, *open-source* et gratuite, est disponible sur le web (https://fongwama.github.io/EduPalu/) et pour les téléphones Android (https://play.google.com/store/apps/details?id=com.fcrm.edupalu&hl=fr).

L'application `EduPalu` a terminé 2^e (sur plus de 650 candidats) du **RFI App Chalenge Afrique 2016** [1].

Ce projet s'est malheureusement arrêté suite à la faillite d'un des principaux sponsors (la compagnie aérienne ECAir) en 2017.

4.10 Marquage métabolique au carbone 12

De retour en France en 2016, j'ai intégré, début 2017, l'équipe « Mitochondries, métaux et stress oxydatif » dirigée par Jean-Michel Camadro à l'Institut Jacques Monod, avec pour objectif de mieux comprendre l'analyse de données en protéomique, et d'y contribuer activement.

Les spectromètres de masse à haute résolution utilisés aujourd'hui en protéomique permettent de connaître avec une très grande précision la masse d'une espèce chimique, d'un peptide voire même d'une protéine, en fonction de la technologie utilisée.

Cette grande précision signifie que pour un ion donné, on obtient non pas un seul pic mais plusieurs pics regroupés dans un massif isotopique. Les différents pics d'un massif isotopique résultent de l'abondance naturelle des isotopes stables des éléments C, H, N, O et S (listés table 4.1).

En spectrométrie de masse, l'identification d'un ion nécessite de connaître précisément la masse de l'ion précurseur (c'est-à-dire de l'ion monoisotopique) et d'interpréter finement le schéma de fragmentation MS/MS. À la fin des années 90, le groupe d'Alan Marshall [3-5] a montré que réduire la complexité isotopique d'une protéine améliorait significativement la détermination de l'ion monoisotopique.

Partant de ce principe, Jean-Michel Camadro et son équipe ont cultivé des cellules de la levure *Candida albicans* dans un milieu enrichi en ^{12}C à 99,99% comme seule source de carbone, ce qui a conduit à la synthèse d'acides aminés puis de protéines enrichies en ^{12}C. Il a

1. https://appafrique.rfi.fr/post/159232834345/nos-lauréats-2016

TABLE 4.1 – Abondance relative des principaux éléments présents dans un peptide.[2]

Élément	Isotope	Mass (Da)	Abondance relative (%)
Hydrogène	^{1}H	1,007825	99,9885
	^{2}H	2,014102	0,0115
Carbone	^{12}C	12,000000	98,930
	^{13}C	13,003355	1,070
Azote	^{14}N	14,003074	99,632
	^{15}N	15,000109	0,368
Oxygène	^{16}O	15,994915	99,757
	^{17}O	16,999132	0,038
	^{18}O	17,999160	0,205
Soufre	^{32}S	31,972071	94,930
	^{33}S	32,971458	0,760
	^{34}S	33,967867	4,290
	^{36}S	35,967081	0,020

ensuite analysé ces protéines par spectrométrie de masse et a observé une augmentation significative de l'intensité de l'ion monoisotopique dans le massif isotopique de chaque peptide. À titre d'illustration, les massifs isotopiques du peptide VGEVFINYIQRQNELFQGKLAYLIIDTCLSIVRPNDSKPLDNR dans les conditions normales (NC) et enrichie ^{12}C (12C) sont représentés dans la Figure 4.7. L'intensité normalisée du pic monoisotopique (M_0) est inférieure à 0,1 en condition NC. Cette intensité monte à 0,6 dans la condition 12C. L'enrichissement en ^{12}C « déplace » la distribution du massif isotopique vers les premiers isotopologues.

Lors de l'analyse par spectrométrie de masse, l'apport de ^{12}C augmente significativement l'intensité des ions monoisotopiques (comme illustré sur la figure 4.7), améliorant ainsi les scores d'identification des peptides et des protéines ainsi marqués.

Cette nouvelle méthode appelée *Simple Light Isotope Metabolic Labeling* ou *SLIM-Labeling* a été publié en 2017 [7].

Toutefois, de nombreux organismes présentent des auxotrophies pour des acides aminés, c'est-à-dire qu'ils ne peuvent pas les synthétiser. Ces acides aminées sont alors essentiels et doivent être apportés dans le milieu de culture. Ils ne peuvent donc pas être enrichis en ^{12}C avec la méthode de marquage *SLIM-Labeling*. Malheureusement, de tels acides aminés enrichis en ^{12}C ne sont pas disponibles commercialement à l'heure actuelle. En effet, le glucose enrichi en ^{12}C, lui-même, n'est pas un produit courant de laboratoire et son coût reste élevé (environ 200 € pour 1 g).

Dans ce contexte, j'ai supervisé le travail de Lilian Yang-Crosson lors d'un stage de Master 1. L'objectif du stage était de développer une méthodologie pour calculer les intensités théoriques normalisées M_0 et M_1 des deux premiers isotopologues, à partir de séquences de peptides, en conditions normale (NC) et enrichie en ^{12}C (12C), avec des acides aminés potentiellement non marqués en ^{12}C.

Si on considère un peptide de formule $C_{n(C)}H_{n(H)}O_{n(O)}N_{n(N)}S_{n(S)}$, où n(C), n(H), n(O) et n(S) représentent respectivement le nombre d'atomes de carbone, hydrogène, oxygène, azote et soufre.

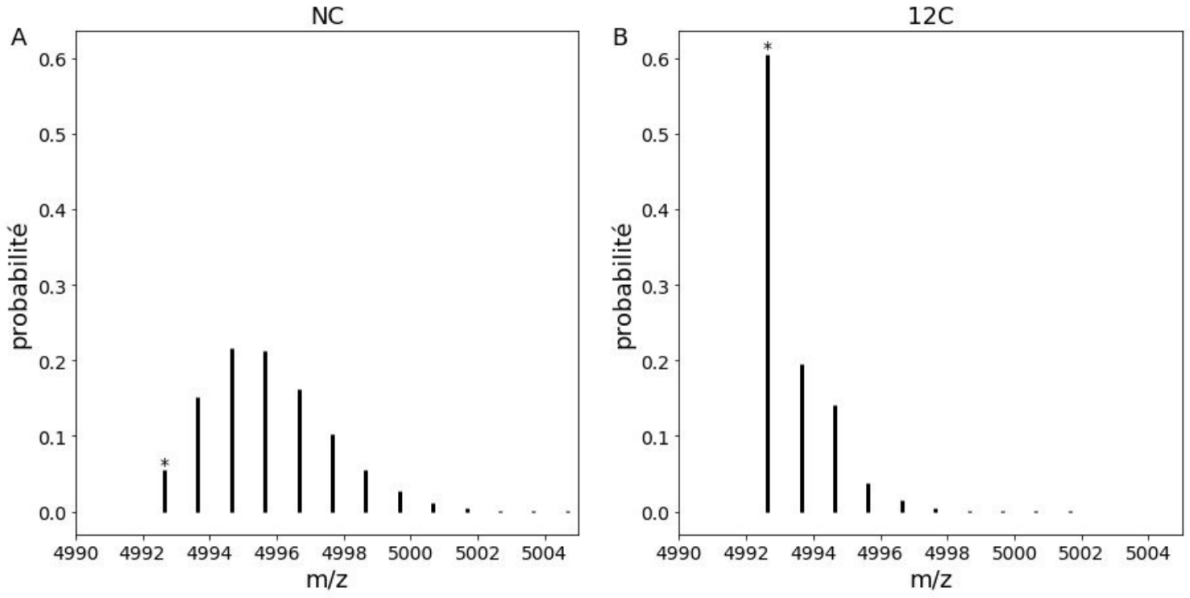


FIGURE 4.7 – Massifs isotopiques théoriques du peptide VGEVFINYIQRQNELFQGKLAYLIIDTCLSIVRPNDSKPLDNR de composition $C_{224}H_{359}O_{66}N_{61}S_1$. Les intensités ont été calculées avec le logiciel MIDAs [6]. Le pic monoisotopique (M_0) est signalé par le symbole $*$. (A) Peptide en condition « Carbon normale » (NC). (B) Peptide en condition « Carbone 12 » (12C).

Pour un tel peptide, l'intensité normalisée de l'ion monoisotopique (M_0) est :

$$M_0 = a(^{12}\mathrm{C})^{n(\mathrm{C})} \times a(^{1}\mathrm{H})^{n(\mathrm{H})} \times a(^{16}\mathrm{O})^{n(\mathrm{O})} \times a(^{14}\mathrm{N})^{n(\mathrm{N})} \times a(^{32}\mathrm{S})^{n(\mathrm{S})}, \tag{4.1}$$

où $a(^{y}\mathrm{X})$ est l'abondance naturelle de l'isotope $^{y}\mathrm{X}$.

Après une expansion polynomiale, l'intensité du 2^e isotopologue (M_1) est :

$$\begin{aligned} M_1 = {} & n(\mathrm{C}) \times a(^{12}\mathrm{C})^{n(\mathrm{C})-1} \times a(^{13}\mathrm{C}) \times a(^{1}\mathrm{H})^{n(\mathrm{H})} \times a(^{16}\mathrm{O})^{n(\mathrm{O})} \times a(^{14}\mathrm{N})^{n(\mathrm{N})} \times a(^{32}\mathrm{S})^{n(\mathrm{S})} \\ & + n(\mathrm{H}) \times a(^{12}\mathrm{C})^{n(\mathrm{C})} \times a(^{1}\mathrm{H})^{n(\mathrm{H})-1} \times a(^{2}\mathrm{H}) \times a(^{16}\mathrm{O})^{n(\mathrm{O})} \times a(^{14}\mathrm{N})^{n(\mathrm{N})} \times a(^{32}\mathrm{S})^{n(\mathrm{S})} \\ & + n(\mathrm{O}) \times a(^{12}\mathrm{C})^{n(\mathrm{C})} \times a(^{1}\mathrm{H})^{n(\mathrm{H})} \times a(^{16}\mathrm{O})^{n(\mathrm{O})-1} \times a(^{17}\mathrm{O}) \times a(^{14}\mathrm{N})^{n(\mathrm{N})} \times a(^{32}\mathrm{S})^{n(\mathrm{S})} \\ & + n(\mathrm{N}) \times a(^{12}\mathrm{C})^{n(\mathrm{C})} \times a(^{1}\mathrm{H})^{n(\mathrm{H})} \times a(^{16}\mathrm{O})^{n(\mathrm{O})} \times a(^{14}\mathrm{N})^{n(\mathrm{N})-1} \times a(^{15}\mathrm{N}) \times a(^{32}\mathrm{S})^{n(\mathrm{S})} \\ & + n(\mathrm{S}) \times a(^{12}\mathrm{C})^{n(\mathrm{C})} \times a(^{1}\mathrm{H})^{n(\mathrm{H})} \times a(^{16}\mathrm{O})^{n(\mathrm{O})} \times a(^{14}\mathrm{N})^{n(\mathrm{N})} \times a(^{32}\mathrm{S})^{n(\mathrm{S})-1} \times a(^{33}\mathrm{S}) \end{aligned} \tag{4.2}$$

Mais dans le cas d'un enrichissement en ^{12}C et d'une auxotrophie (comme par exemple la souche BY4742 de *Saccharomyces cerevisiae* qui ne peut synthétiser les acides aminés histidine, leucine et lysine), les protéines produites par la cellule sont constituées d'acides aminés synthétisés à partir du glucose ^{12}C et d'acides aminés essentiels apportés dans le milieu de culture. Ces protéines seront donc constituées de deux types d'atomes de carbone différents : ceux apportés par le glucose ^{12}C avec une abondance ^{12}C de 99,99% et ceux provenant des acides aminés essentiels avec une abondance naturelle du ^{12}C à 98,93% (et 1,07% ^{13}C). Il convient donc de différencier ces deux types de carbones.

Pour cela, considérons un peptide de séquence ZZZZZZZ où Z peut être n'importe quel acide aminé. Si des acides aminés ne peuvent pas être marqués en ^{12}C (comme dans le cas d'une

auxotrophie), on peut séparer (virtuellement) le peptide en acides aminés marqués (l) et non marqués (nl). Par exemple :

$\mathrm{ZZZZZZZ} \rightarrow \mathrm{Z}^{l}\mathrm{Z}^{l}\mathrm{Z}^{l}\mathrm{Z}^{l} + \mathrm{Z}^{nl}\mathrm{Z}^{nl}\mathrm{Z}^{nl}$

La composition chimique du sous-peptide $\mathrm{Z}^{l}\mathrm{Z}^{l}\mathrm{Z}^{l}\mathrm{Z}^{l}$ est $\mathrm{C}_{n(\mathrm{C},l)}\mathrm{H}_{n(\mathrm{H},l)}\mathrm{O}_{n(\mathrm{O},l)}\mathrm{N}_{n(\mathrm{N},l)}\mathrm{S}_{n(\mathrm{S},l)}$ et celle de $\mathrm{Z}^{nl}\mathrm{Z}^{nl}\mathrm{Z}^{nl}$ est $\mathrm{C}_{n(\mathrm{C},nl)}\mathrm{H}_{n(\mathrm{H},nl)}\mathrm{O}_{n(\mathrm{O},nl)}\mathrm{N}_{n(\mathrm{N},nl)}\mathrm{S}_{n(\mathrm{S},nl)}$.

Pour un marquage au ^{12}C, les abondances isotopiques des éléments H, O, N et S ne sont pas affectées et sont les mêmes dans les deux conditions. On définit alors $n(\mathrm{H})$, $n(\mathrm{O})$, $n(\mathrm{N})$ et $n(\mathrm{S})$, comme :

$$n(\mathrm{H}) = n(\mathrm{H}, l) + n(\mathrm{H}, nl)$$

$$n(\mathrm{O}) = n(\mathrm{O}, l) + n(\mathrm{O}, nl)$$

$$n(\mathrm{N}) = n(\mathrm{N}, l) + n(\mathrm{N}, nl)$$

$$n(\mathrm{S}) = n(\mathrm{S}, l) + n(\mathrm{S}, nl)$$

Et la composition chimique globale du peptide ZZZZZZZ devient alors :

$\mathrm{C}_{n(\mathrm{C},l)}\mathrm{C}_{n(\mathrm{C},nl)}\mathrm{H}_{n(\mathrm{H})}\mathrm{O}_{n(\mathrm{O})}\mathrm{N}_{n(\mathrm{N})}\mathrm{S}_{n(\mathrm{S})}$

L'intensité normalisée de l'ion monoisotopique M_0 s'écrit alors comme :

$$M_0 = a(^{12}_{l}\mathrm{C})^{n(\mathrm{C},l)} \times a(^{12}_{nl}\mathrm{C})^{n(\mathrm{C},nl)} \times a(^{1}\mathrm{H})^{n(\mathrm{H})} \times a(^{16}\mathrm{O})^{n(\mathrm{O})} \times a(^{14}\mathrm{N})^{n(\mathrm{N})} \times a(^{32}\mathrm{S})^{n(\mathrm{S})}, \quad (4.3)$$

Pour le marquage au ^{12}C dans la méthode *SLIM-Labeling*, les abondances sont $a(^{12}_{l}\mathrm{C}) = 0,9999$ et $a(^{12}_{nl}\mathrm{C}) = 0,9893$.

L'intensité normalisé du 2^e isotopologue M_1 devient alors :

$$\begin{aligned}
M_1 = {} & n(\mathrm{C}, l) \times a(^{12}_{l}\mathrm{C})^{n(\mathrm{C},l)-1} \times a(^{13}_{l}\mathrm{C}) \times a(^{12}_{nl}\mathrm{C})^{n(\mathrm{C},nl)} \times a(^{1}\mathrm{H})^{n(\mathrm{H})} \times a(^{16}\mathrm{O})^{n(\mathrm{O})} \times a(^{14}\mathrm{N})^{n(\mathrm{N})} \times a(^{32}\mathrm{S})^{n(\mathrm{S})} \\
& + n(\mathrm{C}, nl) \times a(^{12}_{l}\mathrm{C})^{n(\mathrm{C},l)} \times a(^{12}_{nl}\mathrm{C})^{n(\mathrm{C},nl)-1} \times a(^{13}_{nl}\mathrm{C}) \times a(^{1}\mathrm{H})^{n(\mathrm{H})} \times a(^{16}\mathrm{O})^{n(\mathrm{O})} \times a(^{14}\mathrm{N})^{n(\mathrm{N})} \times a(^{32}\mathrm{S})^{n(\mathrm{S})} \\
& + n(\mathrm{H}) \times a(^{12}_{l}\mathrm{C})^{n(\mathrm{C},l)} \times a(^{12}_{nl}\mathrm{C})^{n(\mathrm{C},nl)} \times a(^{1}\mathrm{H})^{n(\mathrm{H})-1} \times a(^{2}\mathrm{H}) \times a(^{16}\mathrm{O})^{n(\mathrm{O})} \times a(^{14}\mathrm{N})^{n(\mathrm{N})} \times a(^{32}\mathrm{S})^{n(\mathrm{S})} \\
& + n(\mathrm{O}) \times a(^{12}_{l}\mathrm{C})^{n(\mathrm{C},l)} \times a(^{12}_{nl}\mathrm{C})^{n(\mathrm{C},nl)} \times a(^{1}\mathrm{H})^{n(\mathrm{H})} \times a(^{16}\mathrm{O})^{n(\mathrm{O})-1} \times a(^{17}\mathrm{O}) \times a(^{14}\mathrm{N})^{n(\mathrm{N})} \times a(^{32}\mathrm{S})^{n(\mathrm{S})} \\
& + n(\mathrm{N}) \times a(^{12}_{l}\mathrm{C})^{n(\mathrm{C},l)} \times a(^{12}_{nl}\mathrm{C})^{n(\mathrm{C},nl)} \times a(^{1}\mathrm{H})^{n(\mathrm{H})} \times a(^{16}\mathrm{O})^{n(\mathrm{O})} \times a(^{14}\mathrm{N})^{n(\mathrm{N})-1} \times a(^{15}\mathrm{N}) \times a(^{32}\mathrm{S})^{n(\mathrm{S})} \\
& + n(\mathrm{S}) \times a(^{12}_{l}\mathrm{C})^{n(\mathrm{C},l)} \times a(^{12}_{nl}\mathrm{C})^{n(\mathrm{C},nl)} \times a(^{1}\mathrm{H})^{n(\mathrm{H})} \times a(^{16}\mathrm{O})^{n(\mathrm{O})} \times a(^{14}\mathrm{N})^{n(\mathrm{N})} \times a(^{32}\mathrm{S})^{n(\mathrm{S})-1} \times a(^{33}\mathrm{S})
\end{aligned} \quad (4.4)$$

Ce travail théorique a contribué au développement et l'extension de la méthodologie *SLIM-Labeling* pour l'analyse quantitative de peptides (publication #26).

Le marquage au ^{12}C permet de quantifier de manière fiable les protéines provenant de deux conditions différentes, en une seule analyse par spectrométrie de masse. Toute l'information utile à la quantification est en effet contenue dans un unique massif isotopique. À partir des valeurs expérimentales et théoriques des intensités des deux premiers isotopologues M_0 et M_1, il est désormais possible d'obtenir le rapport d'abondance entre deux peptides de même séquence mais synthétisés en conditions NC et 12C, pour des organismes présentant ou pas une auxotrophie.

4.11 Intégration de résultats d'expériences multi-omiques

La quantité de données brutes (fichiers fastq de RNA-Seq, fichiers *raw* de protéomique...) produite par les technologies omiques est énorme et focalise beaucoup l'attention [8, 9]. Ces données brutes, une fois analysées produisent des données secondaires, certes beaucoup moins volumineuses mais multiples et variées, qu'il faut alors stocker, agréger, valoriser.

Avec Gaëlle Lelandais, professeure à l'Institut de Biologie Intégrative de la Cellule (I2BC, UMR CNRS 9198) à l'Université Paris-Saclay, nous avons constaté que les données omiques secondaires, dites « expertes », qui sont les produits des analyses des données primaires, ne sont pas toujours valorisées comme elles le pourraient et sont malheureusement parfois « oubliées » lorsque les premiers résultats scientifiques sont publiés.

Nous avons développé un projet d'intégration de données (secondaires) multi-omiques dont le schéma de principe est représenté Figure 4.8. L'idée est de garder une mémoire de ces analyses en les stockant dans un dépôt de données qui peut être publique ou limité à un usage interne au laboratoire. Chaque analyse est décrite par un vocabulaire contrôlé et liée aux données brutes générées, au protocole expérimental employé et bien sur à la technique d'analyse haut débit utilisée. L'intégration de données externes, par exemple provenant de publications scientifiques est possible pour peu que le processus de génération et d'analyse des données haut débit soit correctement décrit.

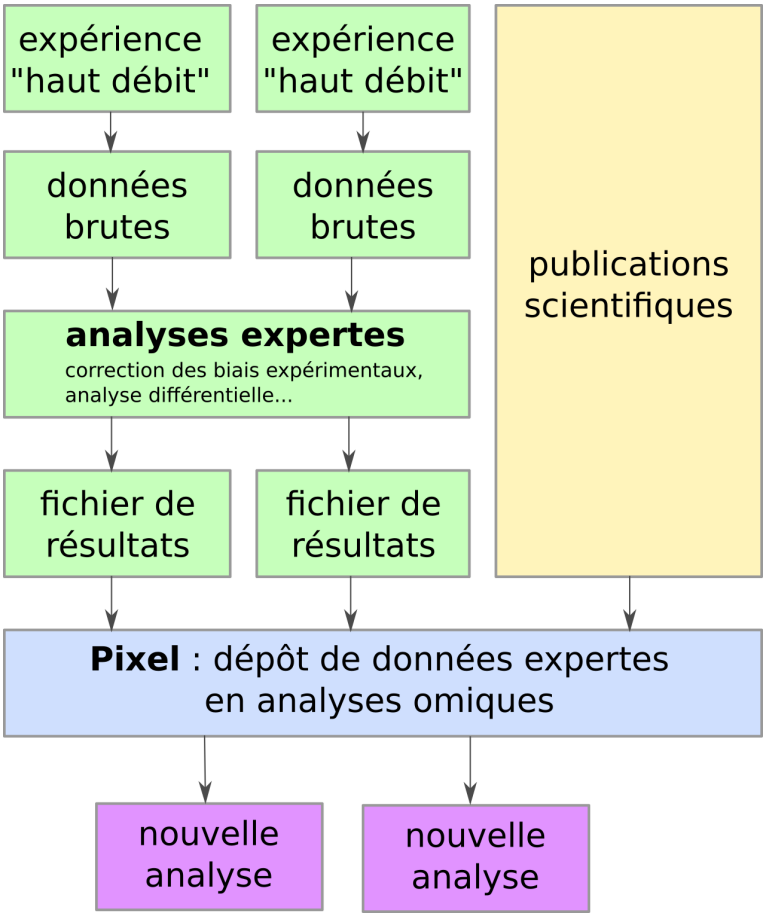


FIGURE 4.8 – Schéma de fonctionnement de Pixel.

L'avantage d'une telle solution est qu'il est ensuite possible, très simplement, d'interroger la base de données et de retrouver les analyses qui ont porté sur telle protéine ou tel gène, qu'elles soient d'origines génomiques ou protéomiques. La combinaison de données d'expériences différentes, venant même de problématiques scientifiques différentes peut répondre à de nouvelles questions scientifiques, tout en valorisant le patrimoine informationnel et scientifique des laboratoires.

Ce projet a été développé par Thomas Denecker pendant sa thèse, en collaboration avec un partenaire privé, TailorDev, qui a mis à disposition 2 développeurs. L'application web produite, `Pixel`, est disponible sous licence *open-source*[2]. `Pixel` a aussi été valorisé sous la forme de la publication #24 dont je suis co-dernier auteur.

2. https://github.com/Candihub/pixel

Partie 5

Cinq activités les plus significatives comme *principal investigator*

À l'heure où les ressources tant financières qu'humaines se font rares, où l'interdisciplinarité, le travail en équipe et l'intelligence collective sont de mises, le terme « ***principal*** *investigator* » semble peu adapté. Je m'essaye néanmoins à cet exercice en proposant ici plusieurs projets dont ma constribution a été déterminante mais certainement pas suffisante.

Plasma (2018–)

Avec Claire Vandiedonck et Sandrine Caburet, nous avons créé le projet « Plateforme d'eLearning pour l'Analyse de données Scientifiques MAssives » (PLASMA). Initié en 2018, ce projet est lauréat des Trophées franciliens de l'innovation numérique dans le supérieur (EdTech2018) et bénéficie d'un soutien financier de l'Idex de l'Université de Paris, de l'EUR GENE et l'UFR SDV pour un budget total de 160 k€.

Ce projet a pour objectif de proposer aux étudiants une plateforme en ligne d'apprentissage pour l'analyse de données omiques (RNA-Seq notamment). Cette plateforme est construite autour d'un serveur Jupyter Hub avec la possibilité de déployer à façon des environnements d'analyses.

L'utilisation des *notebooks* Jupyter [10], véritables cahiers électroniques d'analyses, possède de nombreuses vertus pédagogiques [11]. Mais cette approche est aussi de plus en plus utilisée en recherche, notamment pour documenter des procédures d'analyses [12, 13] et visualiser les résultats obtenus [14, 15].

Pixel (2018–2019)

Développement d'une application d'intégration de résultats d'expériences multi-omiques (`Pixel`). Interaction avec un partenaire privé (TailorDev). Coordination scientifique avec Gaëlle Lelandais.

PBxplore (2015–2017)

Production d'une application pour l'assignation de blocs protéiques à partir de structures tridimensionnelles de protéines (`PBxplore`). Supervision technique du projet. Coordination scientifique avec Alexandre de Brevern.

Technologies numériques contre le paludisme (2015–2016)

Production deux applications mobiles pour le calcul de la densité parasitaire (`DensiPara`) et l'éducation sur le paludisme (`EduPalu`) en République du Congo. Montage du financement du projet. Supervision scientifique et technique d'une équipe de développement logiciel. Coordination avec une ONG (FCRM) et des partenaires privés (Total E&P Congo, Skepsos, ECAir).

XVIIe congrès du Groupe de Graphisme et Modélisation Moléculaire (2011)

Avec Sophie Sacquin-Mora, Marc Baaden et Florent Barbaut, nous avons organisé le XVIIe congrès du Groupe de Graphisme et Modélisation Moléculaire (GGMM). Ce congrès a eu lieu du lundi 30 mai au mercredi 1er juin 2011.

Ce congrès a rassemblé 120 personnes. Au total, 27 communications orales et 62 communications par poster ont été présentées. Le site de la conférence est toujours en ligne : https://ggmm2011.wordpress.com/.

Partie 6

Capacité à organiser des activités de recherche et encadrer des étudiants

Jusqu'à présent, j'ai encadré les activités de recherche de 2 étudiants en thèse, 4 étudiants de Master 2, 4 étudiants de Master 1, 3 étudiants de Licence 3 et 8 développeurs.

Organisé, rigoureux et bienveillant, j'ai valorisé le travail des personnes que j'ai encadrées sous la forme d'articles scientifiques et/ou de productions logicielles. Pour rappel, voici quelques projets pour lesquels mes capacités d'organisation ont été utiles :

bSLIM (2019–2021)

— Encadrement d'un étudiant de Master 1 (Lilian Yang-Crosson).
— Développement méthodologique pour la prise en compte de l'auxotrophie dans le marquage *SLIM-labeling*.
— Publication d'un article scientifique (#26).
— Supervision technique du projet. Coordination scientifique avec Jean-Michel Camadro.

Pixel (2017–2019)

— Supervision de 2 développeurs de la société TailorDev avec Gaëlle Lelandais.
— Production d'une application web (`Pixel`) et publication d'un article scientifique (#24).
— Suivi régulier du projet, notamment sur GitHub.

PBxplore (2015–2017)

— Encadrement d'un ingénieur d'étude (Hubert Santuz) et d'un post-doc (Jonathan Barnoud).
— Production d'une application (`PBxplore`) et publication d'un article scientifique (#22).
— Supervision technique du projet. Coordination scientifique avec Alexandre de Brevern.

Technologies numériques pour lutter contre le paludisme (2015–2016)

— Encadrement des 6 développeurs bénévoles de Fongwama, plateforme congolaise de développement libre.
— Production de deux applications mobiles `DensiPara` et `EduPalu`
— Supervision technique et scientifique du travail des développeurs. Coordination scientifique avec Félix Koukouikila-Koussounda et Francine Ntoumi (FCRM). Gestion de projet développement logiciel libre avec Clémentine Langlois (Skepsos).

Prévalence du VIH en République du Congo (2014–2016)

— Encadrement d'une doctorante (Laure Ghoma Linguissi).
— Travail sur la prévalence du VIH en République du Congo. Publication de 2 articles scientifiques (#18, #20).
— Supervision directe de Laure. Coordination scientifique avec Francine Ntoumi (FCRM).

Partie 7

Projets et perspectives de recherche

Résumé : De nombreux fichiers résultant d'expériences scientifiques en biologie s'accumulent dans des entrepôts de données. La prise de conscience autour de la science ouverte, des données FAIR et de la science reproductible constitue une opportunité pour collecter, analyser, visualiser et valoriser des données de la recherche déjà existantes. Depuis la biologie structurale jusqu'à la biologie haut débit dite « omique », le recyclage de ces données pour produire de nouvelles informations, voire de nouvelles connaissances est un véritable défi tant technique, méthodologique que scientifique.

7.1 Projets scientifiques : vers plus de science ouverte ?

La science ouverte repose sur 3 fondamentaux : l'accès libre aux publications scientifiques (*open access*), la diffusion des données de la recherche (*open data*) et la publication des logiciels scientifiques sous licences libres (*open source*).

Au delà de considérations philosophiques ou réglementaires, avec notamment un effort marqué de l'Europe sur les aspects *open access* et *open data* avec le programme Horizon 2020 (H2020) [16], la science ouverte est aussi un moyen de tendre vers la reproductibilité des analyses, des résultats, et plus globalement de la démarche scientifique. Du fait de la nature virtuelle des artefacts qu'elle manipule, la bioinformatique devrait pouvoir tendre vers cet objectif de reproductibilité. En théorie tout du moins [17-19].

Pour en revenir à la science ouverte, la publication des principes FAIR en 2016 [20] a fait passer l'*open data* d'un concept à un ensemble de règles un peu plus pragmatiques visant à accompagner sa mise en œuvre [1]. Les données de la recherche « FAIR » devraient donc être :

- Trouvables (*Findable*), c'est-à-dire associées à des métadonnées (description, mot-clés) indexées et à un identifiant pérenne (DOI par exemple).
- Accessibles, avec un accès aux données et métadonnées par des protocoles de communications, et éventuellement d'authentification, standards (ftp, http, API documentée).
- Interopérables, avec notamment l'utilisation de formats de données ouverts et documentés (FASTQ, mzML, CSV...). Les métadonnées sont représentées dans un vocabulaire contrôlé.
- Réutilisables, avec des métadonnées riches et normalisées, et une licence associée favorisant la réutilisation, de type Creative Commons Attribution (CC BY) ou Attribution - Partage à l'identique (CC BY-SA), par exemple.

Un des principaux objectifs des principes FAIR est la réutilisation des données, et ce, de façon la plus automatisée possible. D'où l'importance de formats et de protocoles standards et ouverts, et de métadonnées normalisées.

Mais pour que les données ouvertes soient réutilisables, il faut déjà qu'elles soient trouvables et accessibles, donc stockées sur des serveurs accessibles et pérennes. Des entrepôts de données, dont la plupart existaient d'ailleurs bien avant la publication des principes FAIR, sont disponibles à cet effet. On peut citer des entrepôts dédiés comme GEO [21, 22], SRA [23] ou ENA [24] pour les données génomiques, PRIDE [25, 26] ou MassIVE [27] pour les données de protéomiques ou des entrepôts plus généralistes comme Zenodo [28] (projet européen dont l'infrastructure de

1. http://www.go-fair.org/fair-principles/

stockage repose sur celle du CERN) ou figshare [29] (qui appartient désormais au groupe privé *Digital Science*).

Que ce soit pour suivre les consignes des éditeurs scientifiques (notamment pour les données brutes de génomiques et protéomiques) ou pour adhérer au mouvement de la science ouverte, les chercheurs, biologistes et bioinformaticiens, déposent de plus en plus leurs données dans des entrepôts ouverts, en respectant autant que possible les principes FAIR.

On peut alors se demander ce qu'il advient de ces données une fois que, valorisées par une ou deux publications scientifiques, elles se retrouvent stockées dans les entrepôts mentionnés précédemment. Le stockage pérenne et en accès libre de ces données est bien sûr indispensable pour garantir la reproductibilité des analyses dont elles sont la source ou le produit.

Mais peut-on utiliser ces données pour répondre à d'autres questions scientifiques ? Peut-on les interconnecter, les comparer avec d'autres données pour créer de nouvelles informations et avec un peu de chance, de nouvelles connaissances ? On entend beaucoup parler des principes FAIR, mais à quel point les données en *open data* sont-elles vraiment réutilisables ?

C'est avec toutes ces questions en tête que j'ai initié plusieurs projets : *MDbay*, *Missing peptidome* et *MinOmics*.

7.2 MDbay

Contrairement à ce qui est disponible pour les données haut débit « omiques », il n'existe pas d'entrepôt dédié pour le stockage et l'archivage des données issues de simulations de dynamique moléculaire. Pour autant, le paramétrage d'une simulation de dynamique moléculaire requiert un véritable savoir faire avec l'optimisation de la structure de départ, le choix du champ de forces et les paramètres de la simulation elle-même (température, pression, thermostat, algorithme de minimisation...). Quant aux trajectoires produites, qui contiennent les positions de milliers, voire millions d'atomes enregistrées plusieurs millions de fois, elles nécessitent d'importants moyens de calculs pour être générées et sont stockées sous forme de fichiers très volumineux qui contiennent énormément d'informations. Ces raisons justifient largement le besoin de valoriser ces données.

La communauté des chercheurs qui modélisent les lipides a d'ailleurs pris conscience de ce besoin, puisque le projet NMRlipids [30] est en train de constituer sur Zenodo une base de données de trajectoires de dynamiques moléculaires pour les lipides [2].

Le projet MDbay a pour objectif d'indexer et d'analyser les données associées à des simulations de dynamique moléculaire, tant les paramètres et les protocoles de ces simulations que les trajectoires produites, et ce indépendamment du système moléculaire modélisé et du moteur de dynamique moléculaire utilisé (Gromacs, NAMD, Amber, Desmond...).

Ce projet est né d'un workshop Bioexcel qui a eu lieu en 2018 à Stockholm intitulé « Sharing Data from Molecular Simulations ». Il est réalisé en collaboration avec Johanna Tiemann chercheuse post-doctorante au *Linderstrøm-Lang Centre for Protein Science* à l'Université de Copenhague, Matthieu Chavent chercheur CNRS à l'Institut de pharmacologie et de biologie structurale (IBPS) de Toulouse et Steven Garcia, ingénieur en cybersécurité chez Booking.com à Amsterdam.

2. https://zenodo.org/communities/nmrlipids

Pour le moment les données de simulations de dynamique moléculaire ont été indexées depuis l'entrepôt de données figshare. Nous sommes en train de développer les connecteurs pour GitHub et Zenodo. À terme, les entrepôts Dryad, OSF et DataVerse seront également indexés.

Dans un premier temps, nous allons nous intéresser aux fichiers de configuration des simulations, notamment les structures initiales et les paramètres de simulation. Ces fichiers sont souvent de petite taille et au format texte ce qui simplifie leur manipulation.

Nous espérons développer une première preuve de concept qui pourrait répondre aux questions suivantes : Où sont principalement stockés les fichiers ? Quels types de fichiers sont majoritairement disponible ? À quelles températures sont réalisées les simulations de dynamique moléculaire ? Combien d'atomes contiennent les systèmes biologiques simulés ? Quels thermostats sont principalement utilisés ?

Au-delà des réponses à ces questions, le projet proposera aussi aux chercheurs la possibilité d'interroger et d'explorer interactivement les données indexées. L'objectif à terme est d'encourager la réutilisation et la valorisation des données de dynamique moléculaire existantes [31].

7.3 Missing peptidome

La protéomique est la technique haut-débit la plus utilisée pour identifier et quantifier les milliers de protéines d'un échantillon biologique.

Dans l'approche *bottom-up* classique, une protéine est digérée en peptides puis analysée dans un spectromètre de masse. L'étape d'identification des peptides repose principalement sur un algorithme de comparaison des spectres de masse expérimentaux avec des spectres théoriques calculés à partir d'un protéome connu. Lors de cette étape, des hypothèses sont faites sur les modifications post-traductionnelles potentiellement présentes sur les peptides et protéines.

Mais dans une expérience de spectrométrie de masse, certains peptides, pourtant présents dans l'échantillon biologique ne sont jamais identifiés, en raison de la technique expérimentale de spectrométrie de masse ou de la procédure d'identification des peptides.

L'objectif du projet *Missing peptidome* est d'identifier et caractériser ces peptides « absents ». Ce projet a été initié en 2020 par Akram Hecini, étudiant en master 2 en bioinformatique.

Nous nous sommes intéressés au protéome de deux organismes modèles : la levure *Saccharomyces cerevisiae* et la drosophile *Drosophila melanogaster*. À partir d'une digestion trypsique de ces deux protéomes, nous avons interrogé le site *Global Proteome Machine Database* (GPMdb) [32] qui compile les identifications de plusieurs millions d'expériences de spectrométrie de masse analysées par le moteur d'identification X !Tandem [33]. Pour ce projet, nous avons volontairement occulté la variabilité biologique des protéomes car les données extraites de GPMdb étaient indépendantes des conditions expérimentales. Mais, il faut toutefois garder à l'esprit que certaines protéines ne sont exprimées que dans des conditions bien particulières.

Pour chaque peptide de chaque organisme, GPMdb propose un *evidence score* (EC) qui évalue la fréquence d'observation de ce peptide dans les expériences de spectrométrie de masse indexées par GPMdb. Ce score prend les valeurs 1, 2, 3 ou 4 et qualifie des peptides jamais (EC 1) ou peu (EC 2) observés jusqu'à des peptides observés fréquemment et avec certitude (EC 4).

La figure 7.1 représente la distribution de la taille (en acides aminés) des peptides de la levure pour chaque valeur du score EC. Les peptides déjà observés (avec un EC de 2, 3 ou 4)

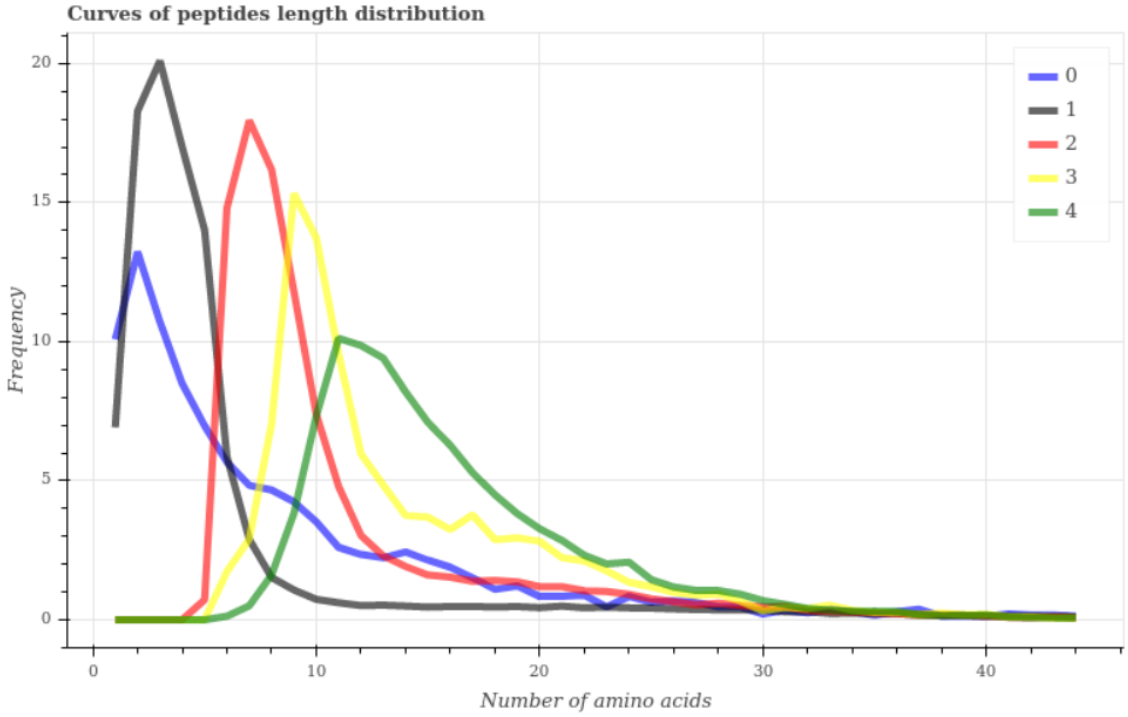


FIGURE 7.1 – *Distribution de la taille des peptides de la levure en fonction du score EC attribué par GPMdb. Le score de 0 a été créé pour représenter les peptides associés à des protéines non référencés dans GPMdb (c'est-à-dire jamais analysées).*

présentent une distribution de taille comprise entre 5 et 35 acides aminés. Cette distribution résulte de la limitation des spectromètres de masse qui autorisent habituellement des ions multi-chargés avec des rapports m/z de 400 à 1800. Inversement, les peptides avec un score EC de 0 ou 1, non répertoriés dans GPMdb, ont des tailles majoritairement petites, à l'exception de certains peptides jamais observés (EC = 0, courbe bleue) qui présentent cependant une taille compatible avec une détection en spectrométrie de masse. Nous avons obtenu des résultats similaires pour la drosophile.

Par la suite, nous n'avons considéré que les peptides ayant une taille comprise entre 5 et 35 acides aminées, c'est-à-dire a priori compatibles avec une détection par un spectromètre de masse. Nous avons séparé ces peptides en deux classes. Ceux qui ne sont pas ou peu observés dans GPMdb (EC = 0, 1 ou 2) et ceux observés régulièrement et avec une bonne confiance (EC = 3 ou 4).

Pour tous les peptides pas ou peu observés, nous les avons recherché à nouveau dans GPMdb mais pour n'importe quel organisme cette fois.

	Levure	Drosophile
Nombre de peptides non trouvés ou peu trouvés chez un autre organisme (EC = 0, 1 ,2)	158 234 (97%)	94,375 (83%)
Nombre de peptides trouvés avec confiance chez un autre organisme (EC = 3, 4)	4 523 (3%)	19 349 (17%)

TABLE 7.1 – Pourcentages de peptides pas ou peu observés chez la levure ou la drosophile mais identifiés dans d'autres organismes.

Les résultats dans le tableau 7.1 indiquent que 3% des peptides pas ou peu observés chez la levure dans GPMdb se retrouvent observés avec confiance dans d'autres organismes. Cette proportion atteint 17% pour la drosophile. Nous avons émis l'hypothèse que des modifications post-traductionnelles, présentes sur certains peptides chez la levure ou la drosophile mais absentes pour ces mêmes peptides chez d'autres organismes rendaient l'identification de ces peptides impossibles chez la levure ou la drosophile. Ces résultats montrent l'importance du choix des modifications post-traductionnelles à prendre en compte lors de l'étape d'identification des peptides. En effet, pour des raisons d'explosion combinatoire, ce ne sont souvent que les modifications post-traductionnelles les plus fréquences (méthylation, acétylation et phosphorylations) qui sont recherchées. Mais ces résultats n'expliquent pas pourquoi des peptides sont pas ou peu identifiés quels que soient les organismes.

Pour répondre à cette question, nous avons créé deux classes de peptides pour la levure et la drosophile ayant une taille de 5 à 35 acides aminés. Les peptides pas ou peu identifiés chez tous les organismes dans GPMdb (EC = 0, 1 ou 2) et ceux bien identifiés (EC = 3 ou 4). À partir de la séquence de ces peptides, nous avons utilisé le logiciel iFeature [34] pour calculer plusieurs propriétés comme la composition en acide aminé, l'ordre des acides aminés, la charge... Nous avons utilisé l'algorithme de classification supervisée par *Random Forest* pour essayer de comprendre quelles propriétés étaient importantes pour expliquer la présence d'un peptide dans GPMdb, donc son analyse en spectrométrie de masse. Que ce soit pour la levure ou la drosophile, les propriétés les plus importantes étaient la charge et la présence de Lysine dans la séquence. Ces propriétés sont liées à la capacité d'un peptide à s'ioniser, notamment en ion positif, ce type d'ion étant majoritairement analysé en spectrométrie de masse. Ces résultats sont cohérents avec ce qu'on peut attendre d'un peptide ionisable en spectrométrie de masse.

Ce projet montre l'intérêt d'analyser de grosses quantités de données annotées pour comprendre finement les propriétés d'un peptide qui gouverne sa détection puis son identification en spectrométrie de masse. Ce premier travail initié par Akram pendant la pandémie de Covid-19 mériterait d'être approfondi. En effet, de nombreuses interrogations subsistes, notamment sur la pérennité de GPMdb car la réactivité du site est parfois très variable. Par ailleurs, nous n'avons pas pu obtenir les paramètres exacts du *workflow* d'analyse qui réanalyse toutes les données expérimentales de spectrométrie de masse pour en indexer les identifications dans GPMdb, notamment les modifications post-traductionnelles recherchées. Une solution complémentaire pour valider ces premiers résultats serait d'interroger la base de données PeptideAtlas [35] qui propose des résultats similaires ou l'entrepôt PRIDE qui offre la possibilité de rechercher directement des séquences de peptides dans les fichiers d'identification déposés par les chercheurs.

7.4 MinOmics

Le projet *MinOmics*, quant à lui, s'intéresse à des données plutôt d'origine génomique.

L'abondance des données haut-débit omiques est aujourd'hui une réalité [8, 9]. Comme mentionné précédemment, le stockage de ces données est assuré par des entrepôts de données spécialisés : GEO, SRA, ENA pour les données génomiques, PRIDE ou MassIVE pour les données protéomiques. Les volumes de données stockées dans ces dépôts sont considérables. Encore une fois, on peut se poser la question du recyclage, au sens réutilisation, de ces données.

Avec Gaëlle Lelandais, je co-dirige la thèse de Thibault Poinsignon sur le projet *MinOmics*. Ce projet a pour objectifs la modélisation de réseaux biologiques, l'analyse de données multi-omiques et leur visualisation à large échelle. Mélina Galopin, maîtresse de conférences à l'Université Paris-Saclay encadre également le travail de Thibault.

Nous avons initié le projet avec la levure *Saccharomyces cerevisiae*, modèle particulièrement étudié en biologie. Au-delà d'une littérature scientifique abondante, *S. cerevisiae* dispose également d'une base de données de référence, très bien annotée, la *Saccharomyces Genome Database* (SGD) [36].

Partant du contenu de cette base de données, nous souhaitons confronter les informations de la SGD avec des données de séquençage haut-débit, notamment de Hi-C, qui ont donné lieu à la reconstruction 3D du génome de *S. cerevisiae* à l'interphase [37].

En résumé, nous avons donc d'une part des données tridimensionnelle de l'organisation spatiale du génome et d'autre part, des informations détaillées sur tous les gènes de *S. cerevisiae*. Cette approche holistique vise à confronter ces données et à explorer si l'organisation spatiale d'un génome peut apporter un éclairage nouveau sur un des modèles les plus étudiés en biologie. Par exemple, nous allons rechercher si les sites de fixation de certains facteurs de transcription seraient organisés spatialement.

Un des livrables de ce projet sera la réalisation d'une interface interactive permettant aux chercheurs d'interroger et d'explorer eux-mêmes ces données. Ce travail de visualisation nous semble fondamental, car la manière de présenter graphiquement les données peut influer fortement sur la représentation puis la compréhension des informations auxquelles le chercheur est confronté.

7.5 Des données aux logiciels scientifiques

Le logiciel scientifique est un objet incontournable de la recherche en bioinformatique car il est l'élément indispensable de toute analyse de données. Le logiciel scientifique est une « donnée » complexe, difficile à appréhender, notamment par sa nature « exécutable » et évolutive, et pour laquelle les principes FAIR ne peuvent pas être calqués simplement [38].

D'ailleurs, dans son projet H2020, l'Europe n'est pas très explicite sur le 3^e pilier de la science ouverte : la publication des logiciels de la recherche sous licence *open source*. Pourtant, si l'*open data* peut être comparée aux ingrédients et l'*open access* à une recette de cuisine, il est curieux d'espérer cuisiner un plat sans disposer des ustensiles adéquats, c'est-à-dire sans les outils et logiciels nécessaires, d'autant plus que les bioinformaticiens ont aussi l'habitude de concevoir leurs propres ustensiles de cuisine !

On se rend compte également que les logiciels scientifiques, sont assez mal référencés, tout du moins en biologie. Ainsi, dans une étude de 2012 [17], les auteurs ont sélectionnés aléatoirement 50 articles scientifiques publiés en 2011 et qui utilisent le logiciel d'alignement de *reads* BWA (lui-même publié en 2009 [39]) dans leurs analyses. Sur ces 50 articles, 62 % n'indiquent ni la version du logiciel ni les paramètres nécessaires à son utilisation. Difficile d'être reproductible dans ce cas.

Enfin, citer (correctement) un logiciel scientifique n'est pas non plus une pratique répandue, comme en témoigne les outils et conseils récents sur le sujet [40, 41].

Ma trajectoire scientifique m'a amené à produire, c'est-à-dire à développer ou co-développer, plusieurs logiciels, du C++ au Python, d'un logiciel de simulation Monte-Carlo à des applications mobiles. Pour être franc, je pense que le logiciel de simulation que j'ai développé en thèse est aujourd'hui inutilisable (code source non versionné et peu documenté). Mais depuis 2012, de nombreuses bonnes pratiques sur le développement logiciel scientifique ont été publiées [42-46]. J'ai également pu me former à la gestion d'environnements logiciels scientifiques et le *packaging* Python. Par conséquent, les outils que j'ai développés plus récemment respectent les bonnes pratiques tant en ingénierie logicielle (gestion de versions, documentation, tests, intégration continue, *packaging*) qu'en ingénierie de la reproductibilité (conteneurisation, *notebooks*...).

C'est dans cette démarche de valorisation du code source scientifique et que je suis devenu ambassadeur du projet Software Heritage (`https://www.softwareheritage.org/`) en 2021. Software Heritage [47] est une organisation à but non lucratif, créée en 2016, dont la mission est de collecter, préserver et partager tous les codes sources des logiciels disponibles publiquement. Software Heritage bénéficie du soutien de l'INRIA, de l'Unesco, du CNRS et d'entreprises privées comme Microsoft, Huawei ou Intel. Software Heritage archive d'ores et déjà le code source du système de guidage d'Apollo 11 comme celui du jeu Quake III. Il archive également des logiciels plus spécifiquement utilisés en bioinformatique, comme le moteur de dynamique moléculaire Gromacs [48, 49], le logiciel d'alignement de reads Bowtie 2 [50] et l'outil de visualisation de réseaux Cytoscape [51]. Software Heritage archive aussi automatiquement l'intégralité des paquets source de Debian, PyPI (paquets Python) et CRAN (paquets R).

La fonctionnalité « Save code now » (`https://archive.softwareheritage.org/save/`) de Software Heritage permet d'archiver quasi-immédiatement n'importe quel dépôt public basé sur les systèmes de contrôle de versions Git, Mercurial ou Subversion. L'intégralité de l'historique de développement est conservée et n'importe quelle version peut ensuite être récupérée. Tous les objets archivés dans Software Heritage bénéficient d'un identifiant persistant appelé SWHID [52]. Le SWHID est un identifiant intrinsèque, c'est-à-dire qu'il est vérifiable indépendamment et ne nécessite par un résolveur intermédiaire comme pour le DOI.

Mes interactions avec Software Heritage consistent essentiellement à développer une interface avec la communauté bioinformatique pour promouvoir :

— L'archivage du code source des outils de bioinformatique, du script au logiciel complexe.

— L'adoption de bonnes pratiques d'archivage avec notamment la mise en place de fichiers de métadonnées lisibles par des humains, comme les fichiers `README`, `LICENSE` ou `AUTHORS` ou plus adaptés aux machines comme le fichier `codemeta.json` (`https://codemeta.github.io/`), par exemple.

7.6 Perspectives à plus long terme

Au-delà de ces projets à court ou moyen terme, je trouve qu'il est difficile de se projeter, tant la science évolue vite et les conditions pour effectuer notre travail de recherche au quotidien se dégradent.

En restant raisonnablement optimiste et en essayant de voir un peu plus loin, je peux déjà dégager de grands principes. Tout d'abord, je suis très attaché à l'idée de science ouverte. Pour citer Bernard de Chartres repris ensuite par Newton, l'activité de recherche scientifique consiste

à « se tenir sur les épaules des géants » [3], mais je ne vois pas comment les générations futures de scientifiques pourront se reposer sur les travaux et les connaissances que nous produisons (devenant à notre tour ces fameux « géants ») sans une nécessaire ouverture et transparence des protocoles, des méthodes, des données produites et des résultats obtenus. Je suis aussi convaincu que le sacro-saint article scientifique n'est pas suffisant pour atteindre cet objectif, car comme le mentionnent Buckheit et Donoho en paraphrasant Claerbout : « *An article about computational science in a scientic publication is not the scholarship itself, it is merely advertising of the scholarship. The actual scholarship is the complete software development environment and the complete set of instructions which generated the figures.* » [53]. Enfin, et comme je l'indiquais en introduction à ce manuscrit, je suis très attiré par le travail aux interfaces et la résolution de problèmes en biologie par des moyens informatiques.

Mes projets à long terme vont donc certainement s'articuler autour de cet axe de travail : tenter de décrire et d'expliquer la biologie avec des données ouvertes et des outils informatiques, eux aussi les plus ouverts possibles. Mon parcours de recherche m'a, jusqu'à présent, conduit vers le développement d'algorithmes, de méthodes ou d'outils pour produire et analyser de nouvelles données. À l'avenir, je souhaite explorer les aspects d'interconnexion et de visualisation de données biologiques existantes pour aller dans le sens du recyclage des données déjà produites, que ce soit des données expérimentales « omiques », des données de simulation, des données de la littérature scientifique ou des données expertes résultant de compilation d'analyses.

3. https://en.wikipedia.org/wiki/Standing_on_the_shoulders_of_giants

Conclusion

Les trois projets décrits précédemment (*MDbay*, *Missing peptidome* et *MinOmics*) abordent la question de l'*open data* par le biais de la réutilisation de données scientifiques et de la valorisation de ces données par la production éventuelle de nouvelles informations. Que ces données soient peu (*MDbay*) ou très structurées (*Missing peptidome* et *MinOmics*), les questions d'indexation, de croisement et de représentation de ces données sont des questions ouvertes et entières.

Ces activités de recherche méritent-elles cependant le terme de *research parasites* comme suggéré par Longo et Drazen dans leur éditorial du *New England Journal of Medicine* en 2016 [54] ? Sans doute pas. La science ouverte et la recherche reproductible nécessitent des données ouvertes et accessibles donc a priori réutilisables. Le recyclage de données existantes propose une approche écologique de l'analyse de données, tant dans l'acception d'écosystème des données (diversité et interaction des données) que dans celle de protection de l'environnement en ne générant pas de nouvelles données à partir de nouvelles expériences. Le logiciel scientifique, enfin, au coeur de la reproductibilité des analyses en bioinformatique, est un objet particulier qui mérite une attention soutenue de part l'imposant écosystème de dépendances logicielles qu'il nécessite pour son fonctionnement. Cette abondance des données, cette sensibilité des logiciels, questionnent la bioinformatique, qui loin d'être morte [55], évolue vers une science (des données) encore plus intégrative.

Le métier d'enseignant-chercheur enfin, dont l'activité est statutairement organisée entre recherche et enseignement, est un métier d'équilibriste, mais aussi un métier d'une extrême richesse, tant sur ses aspects humains, que scientifiques et parfois logistiques. Ma pratique d'enseignant a clairement été utile à mon activité de chercheur et réciproquement, depuis la notion de modèle que j'évoque en première année de licence jusqu'à l'analyse de données omiques en formation continue. Maintenir les deux activités, de recherche et d'enseignement, est difficile, impossible même diront certains. J'ai fait le pari d'essayer de progresser dans ces deux facettes de mon métier. Deux ans après ce que je considère comme mon « Habilitation à Enseigner » (CertifiENS, obtenu en 2019), ces lignes clôturent mon mémoire en vue de défendre une Habilitation à Diriger des Recherches.

Bibliographie

1. De Brevern, A. G., Etchebest, C. & Hazout, S. Bayesian Probabilistic Approach for Predicting Backbone Structures in Terms of Protein Blocks. *Proteins* **41,** 271-287. doi :10.1002/1097-0134(20001115)41:3<271::AID-PROT10>3.0.CO;2-Z (2000).
2. Valkenborg, D., Mertens, I., Lemière, F., Witters, E. & Burzykowski, T. The Isotopic Distribution Conundrum. *Mass Spectrometry Reviews* **31,** 96-109. doi :10.1002/mas.20339 (2012).
3. Marshall, A. G., Senko, M. W., Li, W., Li, M., Dillon, S., Guan, S. & Logan, T. M. Protein Molecular Mass to 1 Da by ^{13} C, ^{15} N Double-Depletion and FT-ICR Mass Spectrometry. *Journal of the American Chemical Society* **119,** 433-434. doi :10.1021/ja9630046 (1997).
4. Shi, S. D., Hendrickson, C. L. & Marshall, A. G. Counting Individual Sulfur Atoms in a Protein by Ultrahigh-Resolution Fourier Transform Ion Cyclotron Resonance Mass Spectrometry : Experimental Resolution of Isotopic Fine Structure in Proteins. *Proceedings of the National Academy of Sciences of the United States of America* **95,** 11532-11537. doi :10.1073/pnas.95.20.11532 (1998).
5. Rodgers, R. P., Blumer, E. N., Hendrickson, C. L. & Marshall, A. G. Stable Isotope Incorporation Triples the Upper Mass Limit for Determination of Elemental Composition by Accurate Mass Measurement. *Journal of the American Society for Mass Spectrometry* **11,** 835-840. doi :10.1016/S1044-0305(00)00158-6 (2000).
6. Alves, G., Ogurtsov, A. Y. & Yu, Y.-K. Molecular Isotopic Distribution Analysis (MIDAs) with Adjustable Mass Accuracy. *Journal of The American Society for Mass Spectrometry* **25,** 57-70. doi :10.1007/s13361-013-0733-7 (2014).
7. Léger, T., Garcia, C., Collomb, L. & Camadro, J.-M. A Simple Light Isotope Metabolic Labeling (SLIM-labeling) Strategy : A Powerful Tool to Address the Dynamics of Proteome Variations *In Vivo*. *Molecular & Cellular Proteomics* **16,** 2017-2031. doi :10.1074/mcp.M117.066936 (2017).
8. Marx, V. Biology : The Big Challenges of Big Data. *Nature* **498,** 255-260. doi :10.1038/498255a (2013).
9. Stephens, Z. D., Lee, S. Y., Faghri, F., Campbell, R. H., Zhai, C., Efron, M. J., Iyer, R., Schatz, M. C., Sinha, S. & Robinson, G. E. Big Data : Astronomical or Genomical ? *PLOS Biology* **13,** e1002195. doi :10.1371/journal.pbio.1002195 (2015).

10. Granger, B. E. & Perez, F. Jupyter : Thinking and Storytelling With Code and Data. *Computing in Science & Engineering* **23,** 7-14. doi :10.1109/MCSE.2021.3059263 (2021).

11. Davies, A., Hooley, F., Causey-Freeman, P., Eleftheriou, I. & Moulton, G. Using Interactive Digital Notebooks for Bioscience and Informatics Education. *PLOS Computational Biology* **16** (éditeur Ouellette, F.) e1008326. doi :10.1371/journal.pcbi.1008326 (2020).

12. Wang, Z. & Ma'ayan, A. An Open RNA-Seq Data Analysis Pipeline Tutorial with an Example of Reprocessing Data from a Recent Zika Virus Study. *F1000Research* **5,** 1574. doi :10.12688/f1000research.9110.1 (2016).

13. Beg, M., Taka, J., Kluyver, T., Konovalov, A., Ragan-Kelley, M., Thiery, N. M. & Fangohr, H. Using Jupyter for Reproducible Scientific Workflows. *Computing in Science & Engineering* **23,** 36-46. doi :10.1109/MCSE.2021.3052101 (2021).

14. Perkel, J. M. Data Visualization Tools Drive Interactivity and Reproducibility in Online Publishing. *Nature* **554,** 133-134. doi :10.1038/d41586-018-01322-9 (2018).

15. Piazentin Ono, J., Freire, J., Silva, C. T., Comba, J. & Gaither, K. Interactive Data Visualization in Jupyter Notebooks. *Computing in Science & Engineering* **23,** 99-106. doi :10.1109/MCSE.2021.3052619 (2021).

16. Burgelman, J.-C., Pascu, C., Szkuta, K., Von Schomberg, R., Karalopoulos, A., Repanas, K. & Schouppe, M. Open Science, Open Data, and Open Scholarship : European Policies to Make Science Fit for the Twenty-First Century. *Frontiers in Big Data* **2,** 43. doi :10.3389/fdata.2019.00043 (2019).

17. Nekrutenko, A. & Taylor, J. Next-Generation Sequencing Data Interpretation : Enhancing Reproducibility and Accessibility. *Nature Reviews Genetics* **13,** 667-672. doi :10.1038/nrg3305 (2012).

18. Sandve, G. K., Nekrutenko, A., Taylor, J. & Hovig, E. Ten Simple Rules for Reproducible Computational Research. *PLoS Computational Biology* **9** (éditeur Bourne, P. E.) e1003285. doi :10.1371/journal.pcbi.1003285 (2013).

19. Kim, Y.-M., Poline, J.-B. & Dumas, G. Experimenting with Reproducibility : A Case Study of Robustness in Bioinformatics. *GigaScience* **7.** doi :10.1093/gigascience/giy077 (2018).

20. Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A., Gray, A. J., Groth, P., Goble, C., Grethe, J. S., Heringa, J., 't Hoen, P. A., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S. J., Martone, M. E., Mons, A., Packer, A. L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J. &

Mons, B. The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Scientific Data* **3,** 160018. doi :10.1038/sdata.2016.18 (2016).

21. Edgar, R. Gene Expression Omnibus : NCBI Gene Expression and Hybridization Array Data Repository. *Nucleic Acids Research* **30,** 207-210. doi :10.1093/nar/30.1.207 (2002).

22. Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., Sherman, P. M., Holko, M., Yefanov, A., Lee, H., Zhang, N., Robertson, C. L., Serova, N., Davis, S. & Soboleva, A. NCBI GEO : Archive for Functional Genomics Data Sets—Update. *Nucleic Acids Research* **41,** D991-D995. doi :10.1093/nar/gks1193 (2012).

23. Leinonen, R., Sugawara, H., Shumway, M. & on behalf of the International Nucleotide Sequence Database Collaboration. The Sequence Read Archive. *Nucleic Acids Research* **39,** D19-D21. doi :10.1093/nar/gkq1019 (Database 2011).

24. Leinonen, R., Akhtar, R., Birney, E., Bower, L., Cerdeno-Tarraga, A., Cheng, Y., Cleland, I., Faruque, N., Goodgame, N., Gibson, R., Hoad, G., Jang, M., Pakseresht, N., Plaister, S., Radhakrishnan, R., Reddy, K., Sobhany, S., Ten Hoopen, P., Vaughan, R., Zalunin, V. & Cochrane, G. The European Nucleotide Archive. *Nucleic Acids Research* **39,** D28-D31. doi :10.1093/nar/gkq967 (Database 2011).

25. Martens, L., Hermjakob, H., Jones, P., Adamski, M., Taylor, C., States, D., Gevaert, K., Vandekerckhove, J. & Apweiler, R. PRIDE : The Proteomics Identifications Database. *PROTEOMICS* **5,** 3537-3545. doi :10.1002/pmic.200401303 (2005).

26. Vizcaíno, J. A., Csordas, A., del-Toro, N., Dianes, J. A., Griss, J., Lavidas, I., Mayer, G., Perez-Riverol, Y., Reisinger, F., Ternent, T., Xu, Q.-W., Wang, R. & Hermjakob, H. 2016 Update of the PRIDE Database and Its Related Tools. *Nucleic Acids Research* **44,** D447-D456. doi :10.1093/nar/gkv1145 (2016).

27. Wang, M., Wang, J., Carver, J., Pullman, B. S., Cha, S. W. & Bandeira, N. Assembling the Community-Scale Discoverable Human Proteome. *Cell Systems* **7,** 412-421.e5. doi :10.1016/j.cels.2018.08.004 (2018).

28. Dillen, M., Groom, Q., Agosti, D. & Nielsen, L. Zenodo, an Archive and Publishing Repository : A Tale of Two Herbarium Specimen Pilot Projects. *Biodiversity Information Science and Standards* **3,** e37080. doi :10.3897/biss.3.37080 (2019).

29. Singh, J. FigShare. *Journal of Pharmacology and Pharmacotherapeutics* **2,** 138. doi :10.4103/0976-500X.81919 (2011).

30. Botan, A., Favela-Rosales, F., Fuchs, P. F. J., Javanainen, M., Kanduč, M., Kulig, W., Lamberg, A., Loison, C., Lyubartsev, A., Miettinen, M. S., Monticelli, L., Määttä, J., Ollila, O. H. S., Retegan, M., Róg, T., Santuz, H. & Tynkkynen, J. Toward Atomistic Resolution Structure of Phosphatidylcholine Headgroup and Glycerol Backbone at Different Ambient Conditions. *The Journal of Physical Chemistry B* **119,** 15075-15088. doi :10.1021/acs.jpcb.5b04878 (2015).

31. Antila, H. S., M. Ferreira, T., Ollila, O. H. S. & Miettinen, M. S. Using Open Data to Rapidly Benchmark Biomolecular Simulations : Phospholipid Conformational Dynamics. *Journal of Chemical Information and Modeling* **61,** 938-949. doi :10.1021/acs.jcim.0c01299 (2021).

32. Craig, R., Cortens, J. P. & Beavis, R. C. Open Source System for Analyzing, Validating, and Storing Protein Identification Data. *Journal of Proteome Research* **3,** 1234-1242. doi :10.1021/pr049882h (2004).

33. Craig, R. & Beavis, R. C. A Method for Reducing the Time Required to Match Protein Sequences with Tandem Mass Spectra. *Rapid Communications in Mass Spectrometry* **17,** 2310-2316. doi :10.1002/rcm.1198 (2003).

34. Chen, Z., Zhao, P., Li, F., Leier, A., Marquez-Lago, T. T., Wang, Y., Webb, G. I., Smith, A. I., Daly, R. J., Chou, K.-C. & Song, J. iFeature : A Python Package and Web Server for Features Extraction and Selection from Protein and Peptide Sequences. *Bioinformatics* **34** (éditeur Valencia, A.) 2499-2502. doi :10.1093/bioinformatics/bty140 (2018).

35. Desiere, F. The PeptideAtlas Project. *Nucleic Acids Research* **34,** D655-D658. doi :10.1093/nar/gkj040 (2006).

36. Cherry, J. SGD : Saccharomyces Genome Database. *Nucleic Acids Research* **26,** 73-79. doi :10.1093/nar/26.1.73 (1998).

37. Duan, Z., Andronescu, M., Schutz, K., McIlwain, S., Kim, Y. J., Lee, C., Shendure, J., Fields, S., Blau, C. A. & Noble, W. S. A Three-Dimensional Model of the Yeast Genome. *Nature* **465,** 363-367. doi :10.1038/nature08973 (2010).

38. Lamprecht, A.-L., Garcia, L., Kuzak, M., Martinez, C., Arcila, R., Martin Del Pico, E., Dominguez Del Angel, V., van de Sandt, S., Ison, J., Martinez, P. A., McQuilton, P., Valencia, A., Harrow, J., Psomopoulos, F., Gelpi, J. L., Chue Hong, N., Goble, C. & Capella-Gutierrez, S. Towards FAIR Principles for Research Software. *Data Science* **3** (éditeurs Groth, P., Groth, P. & Dumontier, M.) 37-59. doi :10.3233/DS-190026 (2020).

39. Li, H. & Durbin, R. Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform. *Bioinformatics* **25,** 1754-1760. doi :10.1093/bioinformatics/btp324 (2009).

40. Cosmo, R. D., Gruenpeter, M. & Zacchiroli, S. Referencing Source Code Artifacts : A Separate Concern in Software Citation. *Computing in Science & Engineering* **22,** 33-43. doi :10.1109/MCSE.2019.2963148 (2020).

41. Katz, D. S., Chue Hong, N. P., Clark, T., Muench, A., Stall, S., Bouquin, D., Cannon, M., Edmunds, S., Faez, T., Feeney, P., Fenner, M., Friedman, M., Grenier, G., Harrison, M., Heber, J., Leary, A., MacCallum, C., Murray, H., Pastrana, E., Perry, K., Schuster, D., Stockhause, M. & Yeston, J. Recognizing the Value of Software : A Software Citation Guide. *F1000Research* **9,** 1257. doi :10.12688/f1000research.26932.2 (2021).

42. Morin, A., Urban, J. & Sliz, P. A Quick Guide to Software Licensing for the Scientist-Programmer. *PLoS Computational Biology* **8** (éditeur Lewitter, F.) e1002598. doi :10.1371/journal.pcbi.1002598 (2012).

43. Prlić, A. & Procter, J. B. Ten Simple Rules for the Open Development of Scientific Software. *PLoS Computational Biology* **8,** e1002802. doi :10.1371/journal.pcbi.1002802 (2012).

44. Taschuk, M. & Wilson, G. Ten Simple Rules for Making Research Software More Robust. *PLOS Computational Biology* **13,** e1005412. doi :10.1371/journal.pcbi.1005412 (2017).

45. Lee, B. D. Ten Simple Rules for Documenting Scientific Software. *PLOS Computational Biology* **14** (éditeur Markel, S.) e1006561. doi :10.1371/journal.pcbi.1006561 (2018).

46. Gruening, B., Sallou, O., Moreno, P., da Veiga Leprevost, F., Ménager, H., Søndergaard, D., Röst, H., Sachsenberg, T., O'Connor, B., Madeira, F., Dominguez Del Angel, V., Crusoe, M. R., Varma, S., Blankenberg, D., Jimenez, R. C., BioContainers Community & Perez-Riverol, Y. Recommendations for the Packaging and Containerizing of Bioinformatics Software. *F1000Research* **7,** 742. doi :10.12688/f1000research.15140.2 (2019).

47. Di Cosmo, R. & Zacchiroli, S. *Software Heritage : Why and How to Preserve Software Source Code* in. Proceedings of the 14th International Conference on Digital Preservation, iPRES 2017 (Japan, 2017).

48. Berendsen, H., van der Spoel, D. & van Drunen, R. GROMACS : A Message-Passing Parallel Molecular Dynamics Implementation. *Computer Physics Communications* **91,** 43-56. doi :10.1016/0010-4655(95)00042-E (1995).

49. Lindahl, E., Hess, B. & van der Spoel, D. GROMACS 3.0 : A Package for Molecular Simulation and Trajectory Analysis. *Journal of Molecular Modeling* **7,** 306-317. doi :10.1007/s008940100045 (2001).

50. Langmead, B. & Salzberg, S. L. Fast Gapped-Read Alignment with Bowtie 2. *Nature Methods* **9,** 357-359. doi :10.1038/nmeth.1923 (2012).

51. Shannon, P. Cytoscape : A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research* **13,** 2498-2504. doi :10.1101/gr.1239303 (2003).

52. Di Cosmo, R. in *Mathematical Software – ICMS 2020* (éditeurs Bigatti, A. M., Carette, J., Davenport, J. H., Joswig, M. & de Wolff, T.) 362-373 (Springer International Publishing, Cham, 2020). isbn : 978-3-030-52199-8 978-3-030-52200-1. doi :10.1007/978-3-030-52200-1_36.

53. Buckheit, J. B. & Donoho, D. L. in *Wavelets and Statistics* (éditeurs Antoniadis, A. & Oppenheim, G.)rédigé par Bickel, P., Diggle, P., Fienberg, S., Krickeberg, K., Olkin, I., Wermuth, N. & Zeger, S., 55-81 (Springer New York, New York, NY, 1995). isbn : 978-0-387-94564-4 978-1-4612-2544-7. doi :10.1007/978-1-4612-2544-7_5.

54. Longo, D. L. & Drazen, J. M. Data Sharing. *New England Journal of Medicine* **374,** 276-277. doi :10.1056/NEJMe1516564 (2016).

55. Bourne, P. E. Is "Bioinformatics" Dead? *PLOS Biology* **19,** e3001165. doi :10.1371/journal.pbio.3001165 (2021).