

# Apprentissage automatique de réseaux d'interactions au sein d'écosystèmes marins

Sujet de thèse de doctorat — Université de Lille — 2023

Directeurs : Cédric LHOSSAINE, Sébastien LEFEBVRE

Co-encadrant : Maxime FOLSCHETTE

## Contexte du sujet

Le fonctionnement des écosystèmes marins et des océans est la résultante de nombreuses interactions entre les espèces qui les composent. Ces interactions déterminent directement ou indirectement les services rendus par ces écosystèmes aux sociétés humaines comme la régulation du climat, l'approvisionnement en produits marins (impactant l'économie), ou encore la régulation de la qualité des eaux côtières (impactant l'écologie). Comprendre ces interactions est donc une clef pour déterminer la trajectoire de ces écosystèmes sous l'influence du changement global. Le phytoplancton (formé d'organismes microscopiques, les microalgues) est un compartiment à la base des réseaux trophiques qui est extrêmement sensible aux variations environnementales pour lesquelles il sert donc d'indicateur. Afin d'en suivre et d'en comprendre l'évolution, des réseaux d'observations sur le long terme avec des fréquences de prélèvement élevées ont ainsi été mis en place depuis les années 1990.

On peut notamment citer le cas de l'observatoire SRN qui, centré sur la Manche et le sud de la Mer du Nord, propose des résultats d'échantillonnages de fréquence bi-hebdomadaire depuis 1992, provenant de 9 stations de prélèvement en mer [SRN, 2017]. Ces prélèvements ont pour but de suivre l'évolution des populations de phytoplancton. D'autres jeux de données sont disponibles ou existent sous embargo, et peuvent concerner d'autres ensembles d'écosystèmes.

Pour chaque prélèvement, les données de l'observatoire SRN contiennent les mesures de facteurs abiotiques (facteurs environnementaux : salinité de l'eau, ensoleillement, taux d'azote...) et biotiques (populations de phytoplancton où chaque espèce est comptée indépendamment). À partir de ces données, il est notamment possible de montrer dans quelle mesure les facteurs environnementaux influent sur le développement de certaines espèces de phytoplancton. Dans de précédents travaux, Sébastien LEFEBVRE et ses collaborateurs ont ainsi pu établir la niche écologique abiotique d'espèces de micro-algues à partir de méthodes statistiques multi-variées [Karasiewicz *et al.*, 2018]. Cependant, ces travaux ont suggéré le rôle prépondérant des interactions biotiques (entre espèces) sans toutefois pouvoir le caractériser, ce qui constitue un verrou intéressant à lever. Ces interactions biotiques peuvent être de plusieurs natures, comme la compétition pour l'accès aux ressources, ou des mécanismes d'allélopathie (entrave à la survie ou à la reproduction d'autres espèces).

# Méthodologie

*Learning From Interpretation Transition* (LFIT)<sup>1</sup> est une approche permettant, à partir de données d'observations, de produire un modèle sous la forme d'un programme logique [Inoue et al., 2014]. Elle s'intègre dans le champ de la programmation logique inductive (*Inductive Logic Programming* ou ILP) dont l'objet est l'apprentissage sous la forme de programmes logiques. Plusieurs itérations de LFIT ont été proposées, afin de traiter des cas d'apprentissage spécifiques à certaines sémantiques (non-déterminisme, asynchronicité, multi-valué...).

GULA, la dernière version de LFIT proposée par Maxime FOLSCHETTE et ses collaborateurs, a pour ambition d'être indépendante de la sémantique et d'apprendre les influences de façon agnostique [Ribeiro et al., 2021]. L'avantage de cette approche est de pouvoir apprendre un programme logique représentant un système dont le mode de mise à jour est inconnu, ce qui est généralement le cas des systèmes réels en biologie : l'apprentissage des influences « seules » devient donc accessible théoriquement, le tout soutenu par des preuves d'optimalité. Le modèle produit par cette approche est par nature explicable, car les prédictions qu'il effectue se basent sur des règles logiques compréhensibles.

En pratique, l'application naïve de GULA à des données réelles bruitées mènerait à un sur-apprentissage (*overfitting*). Cependant, il a été montré qu'une sur-couche heuristique simple permet d'obtenir des modèles de prédiction de bonne qualité sans pour autant perdre leur caractère explicable [Ribeiro et al., 2021]. Par ailleurs, afin d'éviter des temps de calcul trop longs, il sera aussi possible d'utiliser un algorithme glouton produisant un résultat approché du résultat optimal, à condition de s'assurer que cette approximation ne change pas l'interprétation du programme.

Enfin, il sera nécessaire d'interpréter les programmes produits, qui malgré leur caractère explicable peuvent comporter des milliers de règles. La production d'un graphe d'influence, synthétisant les interactions entre espèces de phytoplancton apprises par le programme, sera un premier pas vers la compréhension de ces écosystèmes. Par la suite, il sera pertinent de développer une méthode permettant de produire un réseau de réaction, qui synthétise ces interactions sous une forme semblable à des réactions chimiques avec des paramètres cinétiques. Si tous ces paramètres sont connus, il est possible d'en dériver automatiquement des équations différentielles, permettant de réaliser des simulations temporelles du modèle appris. Cependant, si un seul de ces paramètres ne peut être identifié, alors une telle simulation est impossible, et il deviendra nécessaire d'avoir recours à des méthodes d'interprétation abstraite, qui permettent d'analyser la dynamique du système de façon approchée, en analysant ses principales tendances. De telles méthodes devront être proposées pour permettre d'étudier les variations annuelles des différentes espèces de phytoplancton même en l'absence d'informations cinétiques.

## Déroulement de la thèse

En premier lieu, l'application de LFIT aux données SRN, revêt un intérêt certain. Un travail préliminaire sous la forme d'un stage de master a déjà été effectué, montrant que l'approche est fonctionnelle [Iken et al., 2021]. Le déroulement suivant est proposé pour que la présente thèse de doctorat prenne la suite de ces travaux :

- Prise en main de l'aspect théorique de la méthode d'apprentissage (LFIT), des spécificités des données (SRN) et modèles d'écologie marine, et de l'existant [Iken et al., 2021].
- Application de LFIT aux données. Appréciation des différents algorithmes de LFIT.

---

1 LFIT existe actuellement sous la forme d'une bibliothèque Python libre (licence GPL3) disponible à <https://github.com/Tony-sama/pylfit>.

- Intégration de connaissances biologiques préalables issues de la littérature dans le processus d'apprentissage (température nominale des espèces, sensibilité connue aux paramètres environnementaux...).
- Nettoyage du bruit et amélioration de la prédiction. Une méthode basée sur le front de Pareto a déjà été proposée pour permettre de simplifier les règles logiques produites.
- Des analyses les plus exhaustives possibles devront montrer que les choix d'implémentation (algorithme utilisé, nettoyage...) impactent le modèle produit de façon très limitée. Pour cela, il sera possible de conduire des batteries de tests ou d'envisager une approche par optimisation.
- Extraction d'un graphe d'interaction depuis les programmes logiques afin de rendre le résultat humainement interprétable. À nouveau, plusieurs approches pourront être comparées.
- Extraction d'un réseau de réaction depuis les programmes logiques afin de permettre des simulations si les paramètres cinétiques sont identifiables.
- Développement de méthodes d'interprétation abstraite pour analyser la dynamique d'un réseau de réaction même en l'absence de paramètres cinétiques ; application au réseau produit pour étudier les variations périodiques des populations de phytoplancton.
- À chaque étape, validation des modèles obtenus à l'aide des directeurs de thèse avec confrontation à la littérature existante.
- Comparaison avec d'autres méthodes existantes afin de juger de la validité de ces travaux (random forest, modèle Poisson-lognormal [Chiquet *et al.*, 2021], Hierarchical Modelling of Species Communities [Ovaskainen *et al.*, 2017]...).
- D'autres jeux de données pourront être envisagés par la suite.

L'encadrement de cette thèse sera par nature multidisciplinaire. L'apprentissage et l'analyse des modèles relèvent de l'informatique fondamentale. Les données, modèles et résultats relèvent en revanche de l'écologie. Une discussion continue entre les deux disciplines sera donc nécessaire, permettant de valider la pertinence des connaissances intégrées au processus d'apprentissage et la validité des modèles appris.

Ce travail débouchera sur des publications internationales en intelligence artificielle et en écologie, et permettra des participations à congrès internationaux. De plus, tout développement informatique sera rendu disponible sous une licence libre et pourra faire l'objet d'une intégration à l'existant (notamment dans le cas de LFIT).

## Présentation des encadrants

**Cédric LHOSSAINE** est professeur d'informatique à l'Université de Lille et responsable de l'équipe BioComputing au CRISTAL. Il a encadré plusieurs thèses pluridisciplinaires (médecine) et coordonné plusieurs projets de recherche (notamment : ANR JCJC BioSPACE 2008–2012, PIA ICEBERG 2011–2017, ANR MIGAD depuis 2021). Il est membre du conseil d'administration de l'Equipex REALCAT pour CRISTAL, et est co-créateur et coordinateur du GT CNRS Bioss sur la biologie des systèmes. Ses travaux de recherche concernent notamment la modélisation des réseaux de régulation génétique comme ceux manipulés dans cette thèse, la réduction de modèles complexes et les langages à base de règles.

**Sébastien LEFEBVRE** est professeur d'écologie à l'Université de Lille et fait partie de l'équipe Interest du Laboratoire d'Océanographie et de Géosciences (LOG, UMR 8187) à Wimereux. Il est le directeur de la structure de fédération de la recherche (SFR) « Campus de la mer ». Il s'intéresse à la modélisation des interactions au sein des communautés marines. L'une de ses thématiques de recherche concerne l'écologie du plancton et des micro-algues marines, dont il cherche

notamment à déterminer les interactions pour en comprendre et en prédire la dynamique spatiale et temporelle.

**Maxime FOLSCHETTE** est maître de conférences à Centrale Lille Institut et membre de l'équipe BioComputing. Ses travaux portent depuis son doctorat sur la modélisation de systèmes biologiques, notamment par des modèles discrets, et à l'analyse de leur dynamique grâce à différentes méthodes dont l'interprétation abstraite. En tant que contributeur de LFIT depuis 2017, il s'intéresse aussi à l'apprentissage explicable de modèles discrets à partir de données d'observation.

## Collaborations externes possibles

- Katsumi Inoue (professeur au National Institute of Informatics, Tokyo, Japon) en tant qu'initiateur de l'approche LFIT.
- Tony Ribeiro (chercheur indépendant & postdoc à l'Université de Nantes) en tant que contributeur à LFIT et principal développeur de la bibliothèque Python.
- Ifremer / Laboratoire Environnement Ressources de Boulogne-sur-Mer (responsable : Alain Lefebvre) en tant que pourvoyeur de données, notamment SRN.

## Références bibliographiques

[SRN, 2017] SRN - Regional Observation and Monitoring program for Phytoplankton and Hydrology in the eastern English Channel (2017). SRN dataset - Regional Observation and Monitoring Program for Phytoplankton and Hydrology in the eastern English Channel. 1992-2016. SEANOE.  
<https://doi.org/10.17882/50832>

[Karasiewicz et al., 2018] Stéphane Karasiewicz, Elsa Breton, Alain Lefebvre, Tania Hernández Fariñas, **Sébastien Lefebvre**. Realized Niche Analysis of Phytoplankton Communities Involving HAB: Phaeocystis Spp. as a Case Study. *Harmful Algae* 72, 2018.  
<https://doi.org/10.1016/j.hal.2017.12.005>

[Inoue et al., 2014] Katsumi Inoue, Tony Ribeiro, Chiaki Sakama. Learning from interpretation transition. *Machine Learning* 94, 51–79, 2014. <https://doi.org/10.1007/s10994-013-5353-8>

[Ribeiro et al., 2021] Tony Ribeiro, **Maxime Folschette**, Morgan Magnin, Katsumi Inoue. Learning any memory-less discrete semantics for dynamical systems represented by logic programs. *Machine Learning* 11-12, 2021. <https://doi.org/10.1007/s10994-021-06105-4>  
[ILP Best Paper Award](#) lors de la conférence Inductive Logic Programming (IJCLR'2021).

[Iken et al., 2021] Omar Iken, **Maxime Folschette**, Tony Ribeiro. Automatic Modeling of Dynamical Interactions Within Marine Ecosystems. Poster présenté à la conférence Inductive Logic Programming (ILP), 2021.  
Abstract : <https://hal.archives-ouvertes.fr/hal-03347033>  
Poster : [http://maxime.folschette.fr/Iken\\_IJCLR\\_MAIBioSEM\\_Poster.pdf](http://maxime.folschette.fr/Iken_IJCLR_MAIBioSEM_Poster.pdf)

[Chiquet et al., 2021] Julien Chiquet, Mahendra Mariadassou, Stéphane Robin. The Poisson-Lognormal Model as a Versatile Framework for the Joint Analysis of Species Abundances. *Frontiers in Ecology and Evolution* 9, 2021. <https://doi.org/10.3389/fevo.2021.588292>

[Ovaskainen *et al.*, 2017] Otso Ovaskainen, Gleb Tikhonov, Anna Norberg, Guillaume F. Blanchet, Leo Duan, David Dunson, Tomas Roslin, Nerea Abrego. How to make more out of community data? A conceptual framework and its implementation as models and software. *Ecology Letters* 20, pages 561–576, 2017. <https://doi.org/10.1111/ele.12757>