Métaprogramme DIGIT-BIO

Biologie numérique pour explorer et prédire le vivant

Digital biology to explore and predict living systems

Document Directeur - Version 12 mars 2021

Contributeurs :
Coordination : C. Caranta et H. Monod
Cellule de préfiguration :

- INRAE : Caroline Baroukh (département SPE), Pierre Casadebaig (AgroEcoSystem), Julien Chiquet (MathNum), Pauline Ezanno (SA), Vincent Fromion (MathNum), Santiago Gonzalez Martinez (EcoDiv), Olivier Hamant (BAP), Stéphanie Heux (MICA), Fabien Jourdan (AlimH), Marie-Laure Martin-Magniette (BAP), Tristan Mary-Huard (BAP), Christèle Robert-Granié (GA), Thierry Simonneau (AgroEcoSystem), Masoomeh Taghipoor (PHASE)
- Inria : François Fages (Lifeware), Amedeo Napoli (Orpailleur)
- CNRS : Gabriel Krouk (INSB)

Mots clés : Biologie fondamentale et à visée applicative, changements globaux, systèmes complexes et dynamiques, organisme, multi-échelle, évolution, variabilité, diversité, stochasticité, propriétés émergentes, rétroactions, formalisation, intégration, apprentissage, modélisation, simulation, prédiction, visualisation, gestion, pilotage

1 – CONTEXTE ET ENJEUX

• Enjeux sociétaux

La recherche finalisée en biologie a longtemps été guidée et nourrie par des approches disciplinaires « ciblées » : par exemple, en génétique par la caractérisation et l'étude d'un gène contrôlant un caractère d'intérêt, en toxicologie par l'étude d'effets dose-réponse négligeant les effets cocktails, ou encore en physiologie par des études centrées sur un organe et une fonction. Or, face aux grands enjeux d'aujourd'hui (changement climatique, préservation de la biodiversité, santé globale, sécurité alimentaire, énergie, ressources), il est nécessaire de considérer les systèmes biologiques dans leur globalité, leur diversité et leur complexité. Il s'agit de **développer des approches plus intégratives, permettant d'interpréter, d'explorer et de prédire le comportement de systèmes vivants dans une diversité de configurations qui est hors de portée des modes habituels d'expérimentation ou d'observation.**
Les progrès constants en biologie, tant en termes de connaissances, concepts et outils, et les récents apports de la modélisation ouvrent de réelles opportunités pour aborder ces questions en intégrant les échelles spatiales, du niveau moléculaire jusqu'à l'organisme et à la population. Par leur capacité prédictive, ces nouveaux développements permettent de caractériser la dynamique des systèmes biologiques et donc leur robustesse sur le temps long. Dans ce

contexte, **la biologie numérique[1] constitue une approche prometteuse pour décrypter les grandes fonctions et mécanismes du vivant, pour étudier et prédire le comportement du vivant dans des environnements multiples et fluctuants et également pour gérer et piloter ces systèmes.**

En cohérence avec les orientations du plan stratégique INRAE 2030, les recherches conduites au sein du MP DIGIT-BIO contribueront à :

- La caractérisation des réponses adaptatives des organismes et des populations face au changement climatique et autres moteurs de l'évolution du vivant ;
- La mobilisation de la biodiversité et la re-conception des agroécosystèmes pour la transition agroécologique ;
- Aux études sur la santé des organismes et les interconnexions avec l'environnement et l'alimentation (maladies, pollutions, contaminants, exposome[2]) ;
- L'implémentation et l'optimisation des processus biotechnologiques pour des usages plus efficients et circulaires des ressources.

• Enjeux scientifiques : des évolutions profondes en sciences du vivant

Les évolutions en science des données et les technologies du numérique bouleversent incontestablement la recherche en sciences du vivant. Parmi ces évolutions, l'explosion quantitative et qualitative des données en biologie, associée au développement sans précédent des technologies d'acquisition à toutes les échelles, et aux apports de la science ouverte, permettent d'aborder des questionnements nouveaux et plus complexes sur le fonctionnement du vivant.

De façon connexe, les sciences du vivant font l'objet d'un changement de paradigme théorisé dès la fin du 20ème siècle avec la biologie des systèmes. Les « objets » biologiques sont considérés comme des systèmes dynamiques complexes et évolutifs, dont le comportement global ne peut être déduit simplement des propriétés de leurs composantes. La biologie des systèmes offre un cadre de travail pour comprendre et prédire leur comportement par la mobilisation systématique de méthodes mathématiques et informatiques et d'outils de modélisation, d'inférence (calibration, évaluation) et de simulation. Grâce à la formalisation des interactions et des propriétés particulières des éléments constitutifs des systèmes complexes et de leurs dynamiques, et à la prise en compte d'une plus grande variabilité à différents niveaux d'étude, de multiples propriétés émergentes peuvent être observées et intégrées entre différentes échelles spatiales et temporelles.

Ces approches *in silico* bénéficient aujourd'hui de nouveaux outils et méthodes qu'il est indispensable de développer et s'approprier : intégration massive des données et des connaissances, calcul intensif, calcul distribué, nouveaux modèles et méta-modèles, nouveaux algorithmes d'apprentissage automatique, etc. Ces outils et méthodes combinés à l'explosion qualitative et quantitative des données contribuent notamment à l'essor des approches d'analyse

[1] La *biologie numérique* désigne l'ensemble des approches de recherche en biologie s'appuyant sur des données de toutes natures et sur l'usage intensif des mathématiques et de l'informatique pour les exploiter et les interpréter.

[2] L'exposome correspond à la totalité des expositions des facteurs environnementaux que subit un organisme de sa conception à sa fin de vie, complétant l'effet du génome.

« data-driven » et posent la question de leur potentiel de complémentarité avec les approches « concept-driven » pour l'étude du vivant. Ainsi, en complément des approches analytiques, observationnelles et expérimentales, la mobilisation de méthodes sophistiquées de découverte et représentation des connaissances, de modélisation et simulation, de statistique computationnelle et d'intelligence artificielle[3] occupe une place croissante pour comprendre et prédire les processus biologiques et leur réponse à différentes contraintes, mais aussi pour concevoir et piloter ces systèmes.

• Interdisciplinarité, nouveaux standards et pratiques de recherche

Le développement de la biologie numérique est consubstantiel d'une approche de plus en plus interdisciplinaire de la science et d'une vision plus systémique de la biologie. Pour aborder et relier les différents niveaux d'organisation (de la molécule à la population), la biologie numérique doit mobiliser de façon conjointe les disciplines des sciences du vivant et des sciences de l'environnement (biologie moléculaire et cellulaire, physiologie, génétique, génomique ainsi que des disciplines plus intégratives comme l'agronomie, la zootechnie, l'écologie ou la foresterie), les disciplines des sciences formelles (mathématiques appliquées, statistiques, informatique, sciences de l'ingénieur) ou encore la chimie et la physique,.

L'évolution profonde des pratiques de recherche a été accompagnée d'un accroissement des exigences d'accessibilité aux données et de reproductibilité des résultats. L'intérêt stratégique des données et des développements technologiques associés (data science, intelligence artificielle, …) sont devenus des préoccupations centrales des organismes de recherche. S'il reste important d'accroître le recueil et la production de données de manière « intelligente » (au sens *smart-data*), il est tout aussi indispensable d'augmenter leur accessibilité et leur réutilisation par une diversité d'approches. Ainsi les principes FAIR définissent les objectifs attachés à la gestion des données afin qu'elles soient faciles à trouver, accessibles, interopérables et réutilisables. Tirer le meilleur parti des données amène de nouveaux défis pour la gestion des connaissances (ontologies, extraction et représentation de l'information, intégration de données de natures différentes, formalisation mathématique des hypothèses, etc.).

2 – PROGRAMME

• Objectifs et défis scientifiques interdisciplinaires

L'objectif du MP DIGIT-BIO, qui s'inscrit dans la dynamique de la prospective scientifique interdisciplinaire « Approches prédictives pour la biologie et l'écologie »[4], est de soutenir et développer les recherches visant à **comprendre le fonctionnement et prédire le comportement des systèmes biologiques, anticiper les impacts de différentes contraintes sur ces systèmes, et en raisonner la gestion et disposer de leviers d'action.** À moyen terme,

[3] L'intelligence artificielle recouvre ici, essentiellement, les différentes méthodes d'apprentissage qui lui sont associées (machine, *deep, reinforcemnt learning*, etc.), sans exclure une acception plus large (assimilation de données, raisonnement, contrôle).

[4] DOI : https://search.datacite.org/works/10.15454/1.5783037069682676e12

l'ambition est de développer un petit nombre de projets de suivi *in silico* de systèmes biologiques en s'inspirant et en adaptant le concept de jumeau numérique.

Concernant le périmètre, DIGIT-BIO s'adresse aux comportements des systèmes biologiques, par essence dynamiques et complexes, de **l'échelle de la molécule à celle de l'organisme et de la population[5] dans leurs environnements** (environnements de « proximité » biotique, abiotique, pratiques et modes de gestion). Les échelles plus larges traitant par exemple des interactions entre populations au sein des écosystèmes seront traitées dans d'autres méta-programmes (e.g., BIOSEFAIR, CLIMAE).

DIGIT-BIO s'intéresse aux systèmes biologiques et aux questions de recherche pour lesquels il est a priori essentiel de **prendre en compte plusieurs niveaux d'organisation, d'échelles de temps et d'espace, à intégrer par des méthodes appropriées**. Ainsi, DIGIT-BIO traite des questions sur :

- L'intégration et l'interprétation de masses importantes de données ou de connaissances multi-sources issues d'un niveau d'organisation donné ;
- Le passage de niveaux d'organisation élémentaires à un système plus complexe et le dialogue entre le système et le/les niveaux élémentaires ;
- La modélisation du vivant dans son environnement aux échelles de la cellule à l'organisme et la population ;
- V*ia* la modélisation, le suivi de systèmes biologiques et le cas échéant leur gestion et leur pilotage.

DIGIT-BIO vise à soutenir des recherches ambitieuses en biologie quantitative, intégrative et prédictive. Tout en générant des avancées sur le front des développements méthodologiques, il contribuera à l'élaboration de réponses aux grands défis scientifiques d'INRAE en sciences du vivant, avec une ambition de généricité sur le fonctionnement du vivant tout comme dans les méthodes et outils développés. Il n'établit pas de choix ou de hiérarchie a priori sur les approches méthodologiques à privilégier. Le pari est au contraire d'associer et de faire interagir des biologistes avec des spécialistes de méthodes d'apprentissage (approches « data -driven ») et de modélisation (approches « concept-driven ») pour répondre à des questions sur les systèmes biologiques. Ces approches pourront se confronter parfois pour comparer leurs performances (qualités prédictives, par exemple), mais aussi se compléter et s'hybrider, en particulier dans des contextes où la quantité et/ou la qualité de données est limitante.

DIGIT-BIO mobilise en tant que cadre les **trois composantes clés de la biologie numérique** à savoir l'**intégration de données et de connaissances** scientifiques, l'ensemble des problématiques et méthodes liées aux **changements d'échelles** et la **mise en réseau** de différentes disciplines scientifiques. Dans ce cadre général, DIGIT-BIO porte :

- Des enjeux de recherches cognitive et finalisée en biologie par des approches de biologie numérique, avec une ambition de leadership à l'international, en capitalisant sur nos forces dans toutes les disciplines concernées et avec nos partenaires ;
- Des enjeux de formation, fédération, d'animation, d'organisation et de soutien à une nouvelle communauté scientifique interdisciplinaire avec la volonté (i) de soutenir des transversalités

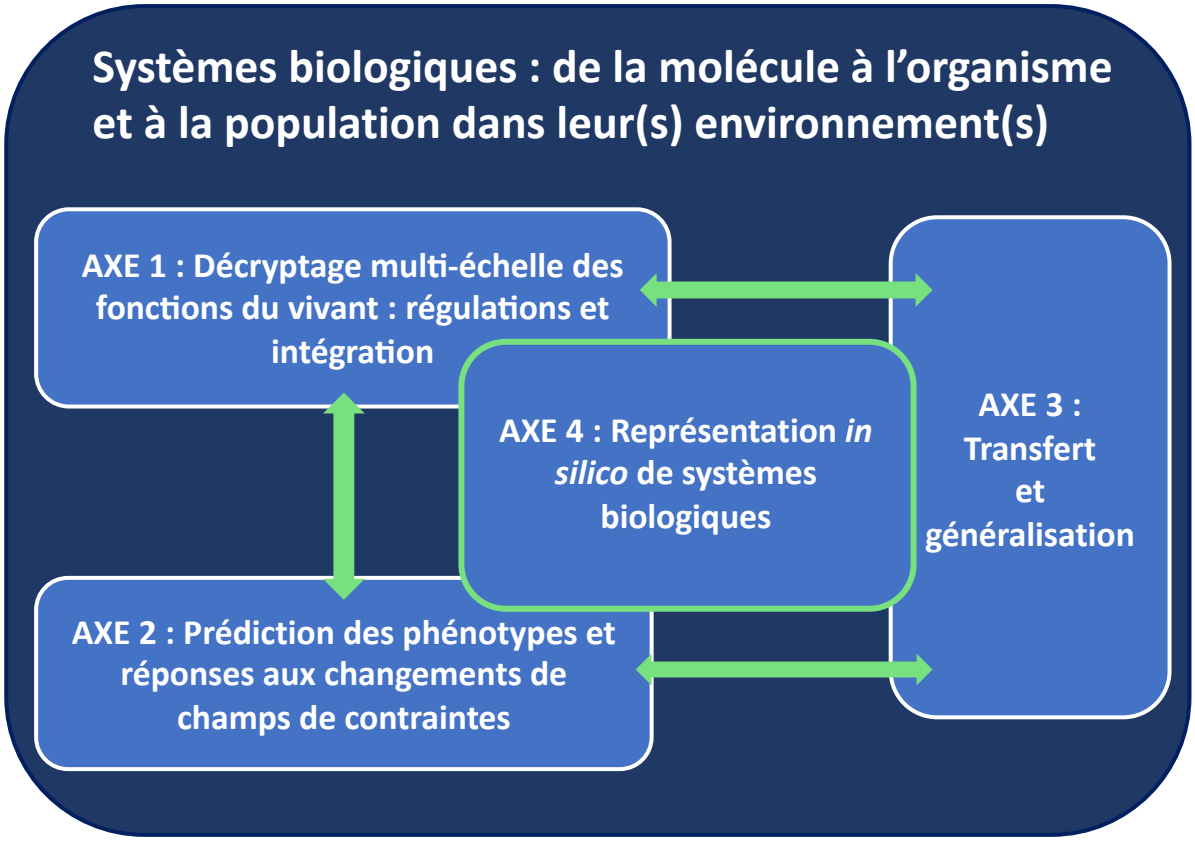
[5] Ensemble des individus d'une même espèce occupant un même lieu

entre les systèmes biologiques étudiés, les échelles d'organisation du vivant et entre thèmes scientifiques différents, (ii) d'encourager la prise de risque et les approches innovantes dans le domaine et (iii) de développer les pratiques favorisant les démarches interdisciplinaires.

Différents leviers seront mobilisés et suivis tout au long de la durée du MP (Cf. partie 7). Ils accompagneront et soutiendront le nouveau sens donné au métier de chercheur et à la recherche dans ses objets d'études (plus complexes), ses qualifications (interdisciplinaires) et ses produits (*open/fair science*) et tireront le meilleur profit de la « mathématisation de la biologie ».

• Axes de programmation

DIGIT-BIO est structuré autour de **quatre axes connexes**. Les axes 1 « Décryptage multi-échelle des fonctions du vivant » et 2 « Prédiction des phénotypes et réponses aux changements de champs de contraintes » seront mis en œuvre dès le démarrage du MP et pourront prendre en compte les questions de transfert et de généralisation (axe 3). L'axe 4, qui correspond à l'ambition à moyen-long terme du MP, mobilisera les connaissances acquises dans les autres axes pour développer la représentation *in silico* de quelques systèmes sélectionnés pour leur intérêt et pour la maturité de leurs modèles.



Axe 1 : Décryptage multi-échelle des fonctions du vivant : régulations et intégration

Cet axe de recherche s'adresse à la **compréhension des processus biologiques, à leurs régulations et à la façon dont ces processus interagissent ou coopèrent**. **Il concerne tous les niveaux d'organisation du vivant jusqu'à l'organisme et à la population.** Il s'agit de décrire, comprendre et modéliser les systèmes biologiques et d'établir les liens causaux au sein et entre les échelles biologiques, en intégrant les effets systémiques comme la stochasticité ou les rétroactions, en tant que déterminants de la dynamique et de l'évolution du système. Comprendre (i) les interactions entre constituants moléculaires dans les structures et fonctions des cellules (microbiennes, végétales, animales, humaines), (ii) comment les cellules croissent, se multiplient, se différencient en réponse aux stimuli de l'environnement ou évoluent vers des fonctionnements dégradés (iii) comment les cellules coopèrent pour former un tissu, un organe, un organisme, (iv) ou encore comprendre les mécanismes adaptatifs des populations et des organismes qui les composent en réponse à différents champs de contrainte et comment ces mécanismes sont mis en place au cours du développement et de l'évolution sont des questions clés. Au-delà de l'étude des processus dans des conditions standardisées, il s'agira également de les caractériser dans une diversité d'environnements (biotique, abiotique, contaminant) et de contextes (e.g., pratiques, modes de gestion, pathologie).

Du point de vue de la biologie numérique, ces défis font appel à la modélisation mathématique et informatique de systèmes dynamiques, au calcul numérique pour la simulation et la calibration, à l'extraction de connaissances et à l'intégration de données hétérogènes pour calibrer et évaluer la qualité des modèles. L'intégration de données issues de différentes sources nécessite de progresser sur les méthodes de représentation des connaissances à l'aide d'ontologies, de vocabulaires contrôlés et de réseaux ainsi que sur les méthodes de spatialisation dans des structures virtuelles aux caractères biophysiques paramétrables (forces d'interactions moléculaires, propriétés biomécaniques,…).. Les méthodes d'apprentissage sont également mobilisées, avec un enjeu d'intelligibilité (fouille de texte basée sur la sémantique) et de recherche de causalité.

Les recherches attendues portent en particulier sur :

- La caractérisation des principes d'organisation, de fonctionnement et d'évolution des génomes, des cellules, des organes et des organismes (en incluant les mécanismes de contrôle de l'expression des gènes) ;
- La construction de réseaux moléculaires sous-tendant les fonctions biologiques, et notamment les réseaux métaboliques multi-échelles ou multi-espèces, en lien avec les processus de régulation ;
- Les liens entre gène(s) et forme(s) en intégrant les contraintes géométriques et physiques et les rétroactions associées, ainsi que ceux entre les régulations moléculaires, le comportement cellulaire et la physiologie des organismes et la façon dont ces liens sont modulés par l'environnement ;
- Les mécanismes sous-jacents aux réponses adaptatives des organismes et des populations et les échelles spatiales et temporelles d'évolution de ces réponses (e.g., potentiel adaptatif lié aux polymorphismes des gènes à la population).

Les défis méthodologiques associés[6] concernent :

- L'extraction et la représentation des connaissances sous des formes permettant la formalisation d'hypothèses et l'intégration dans des modèles (e.g., descriptions des connaissances sous formes de « *linked data* » pour des requêtes complexes, réseaux multiplexés intégrant plusieurs types d'interactions entre molécules et macromolécules, construction d'atlas) ;
- Les relations entre corrélations et causalité, et la recherche de relations de causalité par des méthodes d'apprentissage statistique et automatique sur des données de grande dimension ;
- La modélisation multi-échelle, l'inférence de règles sur les systèmes et l'étude par simulation de systèmes dynamiques complexes décrivant le vivant ;
- La modélisation proprement dite et le couplage de différents modèles (modèles *genome-scale*, modèles mécanistiques, modèles de dynamiques de population et modèles physiques et physico-chimique...) ;
- L'identification de "méta-mécanismes" comme alternative au couplage de modèles entre niveaux d'organisation successifs.

Axe 2 : Prédiction des phénotypes et réponses aux changements de champs de contraintes

L'axe 2 porte sur la **prédiction des phénotypes de la cellule à l'individu et à la population**, de leurs fonctionnalités et des réponses aux changements de champs de contraintes (environnement biotique et abiotique, modes de gestion, pratiques). Comment prendre en compte simultanément l'information biologique et l'information sur l'environnement, ou sur divers champs de contraintes, dans les modèles de prédiction des phénotypes ? Comment prédire le comportement des organismes et des populations face à des modifications de l'environnement ou des pratiques, à différentes échelles temporelles ? Ou encore comment anticiper l'émergence de nouvelles propriétés d'un système en comparaison de celles des entités composant le système ? Ce sont autant de questions clés que DIGIT-BIO propose d'aborder.

Pour la biologie numérique, tout un pan à approfondir porte sur l'élaboration, la comparaison et l'amélioration de méthodes d'apprentissage (statistique, automatique, profond, par renforcement, etc.) issues des développements les plus récents dans ce domaine foisonnant, en les adaptant de façon à intégrer les données multi-sources les plus pertinentes (omiques, capteurs, environnement, données issues de démarches participatives). La modélisation de processus biologiques et physiologiques permettant le développement d'approches prédictives et la simulation de systèmes biologiques complexes constituent également des enjeux majeurs. Il s'agit notamment de tester la robustesse de modèles construits dans l'axe 1, en les soumettant à des conditions fluctuantes externes (environnement) ou interne (physiologie de l'organisme, dynamique cellulaire). L'intérêt du MP inclura donc les approches prédictives reposant sur la modélisation, sur l'assimilation de données et sur le couplage entre les approches dites *data-driven* et *concept-driven*. Il inclura également les dimensions essentielles de la prédiction que sont la planification, la quantification des incertitudes et l'évaluation de la qualité prédictive.

[6] Certains défis méthodologiques sont spécifiques d'un axe du MP alors que d'autres peuvent être partagés entre deux voire plusieurs axes de recherche.

Les recherches porteront sur :

- La prédiction génomique prenant en compte la connaissance des réseaux de gènes, métaboliques et de régulation, ou encore des informations fonctionnelles et environnementales ;
- La prise en compte de l'ensemble des facteurs transmissibles, qu'ils soient génétiques ou non (marques épigénétiques, microbiote, comportement par apprentissage) dans la prédiction du potentiel d'un individu à transmettre ses caractéristiques à sa descendance ou encore des compromis entre fonctions biologiques et leur variabilité ;
- L'application de ces méthodes pour identifier des bio-marqueurs d'état ou de potentiel des organismes (niveau de résistance, niveau de stress, résilience…), ou des cibles génétiques pour l'optimisation de réseaux métaboliques ;
- La prédiction des impacts de scénarios complexes d'exposition chimique et biologique sur la physiologie, de l'échelle de la cellule à celle de l'individu ;
- La mobilisation des informations d'état (phénotype, génome, environnement) et des connaissances sur les processus aux échelles des organismes (et infra) pour prédire le fonctionnement et les trajectoires de populations à différentes échelles spatiales et/ou temporelles.

Les défis méthodologiques associés concernent :

- La sélection des méthodes d'apprentissage les mieux adaptées à un problème donné, la détermination de leurs apports et de leurs limites pour détecter des structures complexes ou des signaux faibles dans les données ;
- La construction et la simulation de modèles intégratifs ayant de bonnes capacités prédictives, et permettant de distinguer les effets de facteurs confondus dans les données ;
- L'accès à des quantités non observables par l'inversion de modèles (e.g., estimation de la valeur de traits non directement mesurables à partir d'une combinaison d'autres traits plus facilement accessibles) ;
- Les questions de qualité et de robustesse de la prédiction (avec les questions sous-jacentes de planification et d'échantillonnage, d'évaluation de l'incertitude, de prise en compte de changements d'environnements).

Axe 3 : Transfert et généralisation

Cet axe de recherche a pour objectifs le **transfert et l'exploitation des résultats et des connaissances acquises sur un niveau d'échelle, un organisme ou un système vers d'autres niveaux d'échelle, organismes ou systèmes** moins étudiés ou partiellement observés, ainsi que le **développement d'approches comparatives plus robustes et valorisant mieux la portée générique des données**. En lien direct avec les recherches qui seront soutenues dans les axes 1 et 2 et en capitalisant sur celles-ci, les objectifs sont de :

- Adapter les approches intégratives et prédictives à des situations d'information partielle et incomplète sur les mécanismes ou processus cibles, à la fois pour tester la généricité de leurs conclusions et pour en étendre les applications ;
- Valider les prédictions construites dans des systèmes modèles en les confrontant à d'autres organismes. Par exemple, dans l'étape d'identification des réseaux de régulation, ou métaboliques l'inférence sur un organisme peut être guidée par la connaissance des réseaux identifiés chez d'autres organismes ;

- Comparer et exploiter des systèmes biologiques différents en vue d'atteindre une plus grande généralisation (par exemple, des organismes avec différents traits d'histoire de vie ou une distribution géographique différenciée) ;
- Développer les méthodes numériques nécessaires à l'extrapolation à l'organisme entier à partir d'observations faites *in vitro* (cellules/organes), un enjeu majeur dans le cadre d'une réduction des études *in vivo* en élevage et en toxicologie.

Les représentations mathématiques ou informatiques de la biologie numérique offrent un cadre propice pour répondre à ces objectifs. Les approches de modélisation et d'apprentissage évoquées dans les axes 1 et 2 sont mobilisables, mais avec des spécificités liées au transfert qui oblige à adapter et évaluer la fiabilité des prédictions alors que l'on a peu de données sur le modèle résultant. Différentes approches se sont développées (apprentissage par transfert, raisonnement et traitement d'analogies, méthodes de projection, découpage et couplage de modules *ad hoc*, génération automatique de modèles). Il s'agit d'en comprendre les propriétés, de les comparer et de les améliorer.

Axe 4 : Représentation *in silico* de systèmes biologiques

La biologie numérique offre la **possibilité d'expérimenter et de suivre *in silico* des systèmes biologiques,** en s'appuyant sur leurs représentations informatiques et possiblement sur leurs mises à jour à partir de données recueillies au cours du temps. Ces possibilités peuvent être exploitées pour améliorer les connaissances et la compréhension, préciser les données supplémentaires restant à acquérir (guider l'expérimentation), anticiper et déterminer comment agir et intervenir sur le système, dans une vision dynamique et interactive que peut favoriser le numérique. La mise en oeuvre de telles approches mobilisera les recherches développées dans les axes 1, 2 et 3. Elle se concentrera sur quelques défis couvrant des domaines et échelles complémentaires pouvant aller par exemple du développement d'un organoïde jusqu'au suivi d'une bioproduction ou d'un élevage, une culture.

Les défis méthodologiques et technologiques porteront en particulier sur les sujets suivants :

- Développer la représentation de quelques systèmes sélectionnés pour leur intérêt et pour la maturité de leurs modèles en s'inspirant du concept de jumeau numérique ;
- Identifier les données réellement nécessaires (smart data) et guider l'expérimentation (e.g. validation expérimentale des résultats) ;
- Explorer le comportement des systèmes modélisés par des méthodes d'expérimentation numérique par exemple pour identifier des assemblages de traits, de variétés/races, d'espèces ou de pratiques ;
- Développer les interfaces homme-modèle, explorer et visualiser l'espace des prédictions, y compris par de la visualisation virtuelle et augmentée ;
- Adapter les processus et piloter les systèmes dans le temps, en réponse à des environnements fluctuants (sélection, biologie de synthèse, traitements de précision, biomarqueurs…) pour développer des stratégies d'intervention. La formalisation de problèmes d'optimisation multi-critères adaptés aux systèmes biologiques (sous-optimalité, résilience, durabilité) est déterminante pour aboutir à des stratégies d'intervention pertinentes.

3 – COMPLEMENTARITES ET VALEUR AJOUTEE

• Par rapport aux schémas stratégiques des départements

DIGIT-BIO aborde des thématiques en sciences du vivant qui font l'objet de défis et/ou d'enjeux structurants pour de nombreux départements aux échelles considérées par DIGIT-BIO: AQUA, MICA, BAP, GA, SPE, SA, AGROECOSYSTEM, PHASE, ECODIV, ALIMH, et TRANSFORM. Le besoin d'acquérir un meilleur accès aux approches mathématiques et numériques pour répondre à des questions en biologie est partagé, que ce soit sur des enjeux d'intelligence artificielle, de modélisation ou de calcul ou des enjeux d'intégration de données. Les sciences formelles portées par le département MATHNUM et de façon croissante par les autres départements ont un rôle central dans DIGIT-BIO. Il constituera un espace favorable pour développer des collaborations interdisciplinaires et fédérer une grande partie des modélisateurs de l'Institut et potentiellement de ses partenaires. Dans tous les cas, les animations sur les méthodes les plus génériques resteront ouvertes à l'ensemble des communautés, afin de favoriser les potentiels couplages entre la diversité des approches de modélisations conduites au sein d'INRAE.

Bien que les sciences humaines et sociales ne soient pas dans le périmètre du MP, il est prévisible que des liens s'établiront avec les départements ACT et ECOSOCIO pour aborder les usages et impacts de certaines recherches, en particulier concernant l'axe 4.

• Disciplines et communautés concernées

DIGIT-BIO mobilise une diversité de disciplines biologiques, de l'environnement et de disciplines formelles. Un objectif majeur est de **décloisonner ces disciplines, les communautés associées et les objets/systèmes d'études** (animaux, plantes, microbes…), pour former une nouvelle communauté autour de la biologie numérique, à l'échelle de l'établissement mais aussi largement ouverte à l'extérieur. Ce décloisonnement est également nécessaire pour le développement de quelques projets emblématiques sur la représentation *in silico* de systèmes biologiques.

Les besoins en recherche et en ingénierie sont omniprésents en biologie numérique. A côté des chercheurs, les ingénieurs informaticiens, statisticiens et mathématiciens des départements font donc partie des communautés concernées. Plusieurs CATIs[7], plateformes et infrastructures de l'institut le seront particulièrement : sur la bioinformatique (e.g., infrastructure BioInfOmics, CATIs BIOS4BIOL, BOUM, EMPREINTE...), la modélisation (e.g., plateformes OpenAlea[8], Virtual Plant, Record, CATIs IUMAN ou SysMics) et le recueil de données (e.g., CATIs GEDEOP, SICPA) même si la production de données *per se* n'est pas au cœur du MP. L'ambition ici est de décloisonner

[7] CATI : Centre Automatique de Traitement de l'Information.
BIOS4BIOL : Bioinformatique et statistiques pour la biologie ; BOUM : Bioinformatics for Omics and metaOmics of Microbes ; EMPREINTE : Molecular PhEnotyping and bIochemical daTa Engineering; IUMAN : Informatisation et Utilisation des Modèles pour les Agroécosystèmes Numériques ; SysMics : System biology for omics ; GEDEOP : gestion des données d'expérimentation ; SICPA : Système d'information et calcul pour le phénotypage animal

[8] OpenAlea : Software environment for plant modelling, http://openalea.gforge.inria.fr/dokuwiki/doku.php

les communautés en analyse des données et modélisations structurées autour de disciplines (e.g., écophysiologie, génomique) ou d'objets.

De façon plus globale, le plan « Données pour la science » et les « Principes de gouvernance des données » en cours d'élaboration à l'échelle de l'institut constitueront un cadre à mobiliser par les recherches du MP pour assurer la protection, la gestion et le partage des données avec toute la vigilance nécessaire sur les enjeux associés. Un point décisif pour le passage à l'échelle de certains projets de biologie numérique concernera l'accès aux e-infrastructures régionales ou nationales dédiées au stockage et au calcul intensif (France Grilles, data centers, super-calculateurs).

• Par rapport aux principaux dispositifs et grands programmes de recherche, au niveau national, européen et international

A l'échelle nationale, DIGIT-BIO porte une ambition de visibilité aux recherches conduites par INRAE et favorisera le développement d'actions communes avec des structures de recherche mises en place pour fédérer des communautés interdisciplinaires autour d'enjeux de **modélisation et d'IA**, notamment les Instituts convergence DATAIA (science des données, intelligence artificielle et société) ou les 3IA (sur les volets environnement et santé). Il partagera des thématiques et bénéficiera d'interactions avec plusieurs dispositifs nationaux mis en place pour fédérer des communautés à l'interface des sciences du vivant et des sciences formelles et en particulier des **infrastructures et e-infrastructures de recherche** au service de ces communautés, dont les ambitions dépassent le cadre strict de la production des données et de la bioinformatique pour recouvrir également la biologie intégrative et la biologie des systèmes : Institut Français de Bioinformatique (IFB), France Génomique, METABOHUB dans le domaine de la métabolomique et fluxomique, Phenome et Liph@SAS dans le domaine du phénotypage végétal et animal, voire THEIA aux échelles les plus larges abordées.

Certains de ces dispositifs et grands programmes de recherche ont une déclinaison à l'échelle européenne. C'est le cas d'ELIXIR sur la bioinformatique ou d'EMPHASIS sur le phénotypage. D'autres grands programmes ou projets abordent des enjeux de biologie numérique dans un périmètre thématique ciblé (e.g., B4EST et FORGENIUS en foresterie, …).

C'est le cas également **d'instituts à l'Europe et l'international** : le CEMB, UPenn (USA) ou le NAIST (Japon) sur la mécano-biologie ; le Weizmann Institute of Science, Israël sur la biologie de synthèse ; l'Institute of Integrative Biology, ETH Zurich sur les interactions entre les organismes et leurs environnements, l'Université d' Uppsala pour les analyses intégrées génotypes-phénotypes, le WUR sur les jumeaux numériques.

Le comité de pilotage stratégique du MP DIGIT-BIO aura un rôle important dans l'analyse des opportunités et la priorisation des interactions avec ces différentes initiatives ou dispositifs, qui pourront être amplifiées, par exemple par la création de Laboratoires Internationaux Associés (LIA), comme cela est déjà initié dans le domaine de la morphogenèse des plantes (LIA-IRL Compumorph CNRS-INRAE-Inria-ENS avec Cambridge, UK), ou de la génétique animale (LIA *Genetic improvement of indian cattle* avec le BAIF, Inde).

• Articulation avec les programmes prioritaires internationaux d'INRAE

Parmi les programmes prioritaires internationaux (PPI) actuels, aucun n'est en prise directe avec DIGIT-BIO. Néanmoins, certains peuvent porter des projets de collaboration internationales rentrant dans les thématiques du MP (e.g., PPI Agroécologie, PPI Adaptation des forêts au changement climatique), par exemple via la question de l'adaptation et de la robustesse face aux fluctuations environnementales. Il sera intéressant dans ce cas d'étudier les possibilités de synergie.

• Interfaces avec d'autres MP

Les articulations seront à travailler et organiser en particulier avec deux MP :

- Issu en partie d'une réflexion partagée lors de la prospective scientifique interdisciplinaire « Approches prédictives pour la biologie et l'écologie », HOLOFLUX partage avec DIGIT-BIO un positionnement amont marqué et un recours important à la science des données et à la modélisation. Les interfaces devront s'organiser sur les questions de prise en compte du microbiome dans la prédiction des phénotypes (dans DIGIT-BIO) et sur les mécanismes d'assemblage et d'interaction au sein des holobiontes, et la maîtrise et le pilotage des flux microbiens (dans HOLOFLUX).
- Dans le domaine de l'alimentaire et de la santé, DIGIT-BIO et le MP Système alimentaire et santé (SYALSA) bénéficieront mutuellement de leurs avancées en particulier sur les questions de toxicologie et d'écotoxicologie prédictives et des analyses multi-échelles et multi-systèmes.

Concernant les autres MP, et de par ses ambitions et son périmètre, DIGIT-BIO apportera des connaissances et de potentiels leviers d'action à mobiliser pour les approches à des échelles plus larges (par exemple avec CLIMAE pour des objectifs d'adaptation et d'atténuation du changement climatique, SANBA et SuMCrop sur des enjeux de santé animale ou végétale, BIOSEFAIR sur la biodiversité et le passage à l'échelle des communautés d'espèces). Il pourra jouer un rôle transversal sur les besoins et l'application de développements méthodologiques pour répondre à certains enjeux de compréhension et de prédiction rencontrés dans les autres MP.

4 – AMBITIONS

• Percées scientifiques potentielles

Les connaissances générées par DIGIT-BIO contribueront à des percées scientifiques sur des fronts de science de la biologie numérique portant sur le décryptage des fonctions du vivant et sur la prédiction des phénotypes, de leurs fonctionnalités et réponses aux changements de contraintes à l'échelle des organismes et des populations. Il s'appuiera pour cela sur des équipes et des projets déjà bien positionnés sur ces sujets et en fera émerger de nouveaux, notamment autour des questions de représentation *in silico* de systèmes biologiques.

Des percées sont aussi attendues dans l'exploitation de la richesse des données produites et collectées par des méthodes souvent associées aujourd'hui à l'intelligence artificielle (IA), mais aussi en modélisation en améliorant notre capacité à prendre en compte les interactions et les

différentes sources de variabilité et d'incertitude. Dans ce cadre, la mise en œuvre de concepts intégratifs tels que celui de jumeau numérique fait partie de l'ambition de DIGIT-BIO comme mentionné dans l'Axe 4.

De façon plus globale, et au-delà de l'ambition d'acquisition de connaissances fondamentales, nous sommes convaincus que répondre aux grands enjeux mondiaux dans les domaines de l'agriculture, de l'alimentation et de l'environnement et à leurs interfaces, nécessite à la fois le développement d'approches systémiques et une connaissance fine des mécanismes et processus sous-tendant les comportements des systèmes biologiques, par essence dynamiques et complexes, ainsi que de capacités de prédiction fortes dans toute la gamme d'échelles abordées dans le MP. Ces connaissances et capacités de prédiction constituent bien un maillon de la vision systémique qu'il est indispensable de développer et en ce sens, DIGIT-BIO est en bonne position pour alimenter la réflexion et soutenir des projets de recherche capables d'apporter des scénarios futurs robustes et des solutions opérationnelles.

• Impacts visés selon les dimensions d'impact ASIRPA

Les impacts attendus des percées scientifiques de DIGIT-BIO concernent plusieurs grands domaines de recherche d'INRAE, notamment :

- La prédiction des phénotypes, de la plasticité phénotypique des individus ainsi que celle de la trajectoire évolutive des populations afin de renforcer leur résilience vis-à-vis des modifications de l'environnement et ainsi proposer des stratégies de sélection et/ou gestion en appui à la transition agroécologique, à l'adaptation au changement climatique et à la préservation de la biodiversité.
- L'intégration de données génomiques (modèles polygéniques, réseaux d'interaction) dans des modèles écologiques afin d'augmenter la précision de prédiction des traits adaptatifs pour les populations jamais testées dans un cadre expérimental.
- La santé des végétaux, des animaux, des humains et de l'environnement. Cela concerne par exemple le développement de biomarqueurs pour des interventions adaptées ou encore la compréhension et la prédiction des effets induits par les contaminations environnementales, y compris alimentaires, sur le vivant, de la cellule à l'individu.
- L'optimisation de processus biotechnologiques, par exemple en proposant des modèles alliant conception de la souche et conduite du procédé, permettant ainsi une meilleure transposition des résultats de la recherche en innovations.

• Enjeux de positionnement national, européen et international

Au niveau national, les enjeux concernent la structuration de la communauté scientifique en biologie numérique dans l'objectif d'établir des transversalités entre les systèmes biologiques étudiés, les échelles d'organisation du vivant et entre thèmes scientifiques. Dans la dynamique d'actions déjà initiées, une attention particulière sera portée au développement d'actions

conjointes inter-organismes en particulier avec l'Inria[9] et également le CNRS[10].

Le domaine de la biologie numérique pour explorer et prédire le vivant est en plein essor au niveau international : nombre croissant de publications, développement de journaux dédiés (e.g., International Journal of Integrative Biology, In Silico Plant, Quantitative Plant Biology, INRAE étant dans certains comités éditoriaux), développement d'instituts dans certains pays (e.g., Intitute of integrative biology, ETH Zurich ; Genome Institute of Singapore, Center for plant integrative biology, UK) ou encore programmes phares (e.g., Digital twin projects du WUR). L'enjeu de positionnement et de compétitivité dans les domaines de recherche d'INRAE est indéniable et la forte dynamique identifiée à INRAE mais aussi à l'échelle nationale avec nos partenaires indique qu'un positionnement européen voire international affirmé serait possible sous réserve d'être proactif et d'entrainer les leaders INRAE et nationaux du domaine.

• Atouts et limites d'INRAE dans le domaine du MP

L'Institut dispose d'un potentiel significatif pour développer les approches intégratives et prédictives en biologie au centre des ambitions de DIGIT-BIO. Il possède des forces conséquentes et reconnues aussi bien dans les disciplines biologiques que mathématiques, informatiques et numériques, ainsi qu'à leur interface avec des équipes leaders dans leur domaine (e.g., pôles de recherche à Lyon, Saclay, Toulouse). L'organisation et les actions programmatiques de l'Institut (départements, métaprogrammes, prospectives scientifiques interdisciplinaires) et certains de ses partenariats (UMR, équipes projets communes avec Inria, LIA, etc.) représentent des atouts pour développer des projets interdisciplinaires ambitieux et au meilleur niveau international. Il est également doté d'équipements, d'infrastructures et e-infrastructures lui permettant de recueillir et organiser des données de qualité, de natures diverses et multi-échelles, indispensables pour mener des recherches originales.

Un premier point de vigilance est celui du risque de dispersion étant donné la diversité et le nombre de thématiques concernées par la biologie numérique dans l'Institut. Le MP devra veiller à ne pas tomber dans ce travers et tenir la capacité de prioriser les projets qu'il soutient en se tenant au périmètre proposé et aux ambitions affichées.

Les autres limites, que le MP devra contribuer à lever, relèvent d'une organisation encore insuffisante autour des enjeux du numérique et de l'informatique scientifique (stockage, calcul) pour « monter à l'échelle ». Ces limites ne pourront être contournées que par des évolutions en interne faisant l'objet du « plan données pour la science » (en cours d'élaboration) et par le renforcement de partenariats autour du numérique. En ce sens, les collaborations européennes et internationales autour des enjeux de la biologie numérique méritent d'être augmentées et mieux structurées (e.g., via Elixir pour la bioinformatique).

[9] Développement d'actions conjointes d'ores et déjà acté (e.g., développement d'équipes projets, défis scientifiques co-construits et soutenus par des moyens des deux instituts)
[10] En lien avec l'appel à projet de la MITI (mission pour les initiatives transverses et interdisciplinaires) du CNRS sur la modélisation du vivant

5 – PARTENARIATS

• National

La liste des partenariats à renforcer sera précisée et priorisée au cours de la vie du MP, mais on peut identifier d'ores et déjà certains partenariats à considérer de façon privilégiée avec :

- Inria, de façon à renforcer les relations qu'INRAE a développées avec cet institut sur les méthodes d'intégration de données, la modélisation, le calcul et la simulation appliquées aux sciences du vivant. L'Inria sera étroitement associé au pilotage du MP à travers sa participation au comité de pilotage stratégique et un soutien conjoint à des actions emblématiques.
- CNRS (INSB, INSMI, INS2I) : DIGIT-BIO permettra de développer les collaborations par des actions concrètes sur certaines questions concernant les grandes fonctions du vivant ainsi que sur la reconnaissance et l'analyse d'images, l'extraction des connaissances, la fouille et la synthèse de données, la modélisation et l'assimilation de données. Des échanges avec la MITI (Mission pour les Initiatives Transverses et Interdisciplinaires) seront à organiser en lien avec les appels à projets 2019 et 2020 sur la « modélisation du vivant » ;
- Les universités et les grandes écoles d'ingénieur partenaires dans des politiques de site, en particulier dans le cadre de l'Université Paris Saclay (AgroParisTech, CentraleSupelec, ENS Paris-Saclay, …) ; le MP sera notamment attentif aux enjeux de formation (pour et par la recherche) et de sensibilisation des étudiants aux thématiques de la biologie numérique ;
- Des organismes expérimentant, collectant et exploitant des données dans nos thématiques (Instituts techniques, ONF), avec lesquels il pourra être pertinent de structurer le partenariat autour des enjeux du MP ;
- Des laboratoires ou instituts positionnés résolument à l'interface entre sciences du vivant et sciences mathématiques, physiques ou informatiques (e.g., Institut Curie, Matières et systèmes complexes, et Centuri (Marseille), centres leaders en biophysique animale).

• Europe et international

Le principe est de s'appuyer sur le MP pour renforcer la présence d'INRAE dans des communautés, réseaux et programmes européens ou internationaux à l'interface sciences du vivant et sciences formelles. Les cibles concerneront plusieurs équipes dans la thématique du MP, et devront être travaillées avec le comité de pilotage stratégique (e.g., ELIXIR et l'EBI sur la bioinformatique et l'intégration de données en biologie).

Nous nous appuierons également sur des réseaux existants : par exemple, le LIA entre l'UMR RDP (Lyon) et le Sainsbury Laboratory (Cambridge) a un rôle d'animateur de la communauté travaillant sur le développement des plantes (workshops annuels réunissant les principaux acteurs internationaux). Une cartographie de ces réseaux/communautés sera réalisée par le MP dans l'objectif de créer des liens entre communautés (e.g., développement X métabolisme). Par ailleurs, le potentiel de développement de réseaux et programmes à l'Europe et à l'international feront partie des critères d'éligibilité des actions qui seront soutenues par DIGIT-BIO.

• Potentiel de partenariat avec les acteurs du monde socio-économique

En lien direct avec le processus de « mise en données » du monde, le développement des technologies d'observation et de calcul ainsi que des nouvelles méthodes de modélisation et d'apprentissage, la biologie numérique constitue un domaine en plein essor pour le développement des partenariats socio-économiques. Si la dynamique est clairement à l'œuvre en marketing, ou encore en santé, elle démarre dans nos domaines. En génétique animale, l'avènement de la sélection génomique dans les années 2000 donne une idée du potentiel d'innovation de rupture apporté par l'introduction de nouvelles méthodes prédictives. Plus récemment, le développement de la phénomique végétale qui embarque le développement et la mise en production de méthodes intégrées pour l'identification et la caractérisation de traits de réponse aux contraintes environnementales utilisables en génétique et modélisation des cultures est un autre exemple[11]. Pour le futur, des représentations *in silico* d'organes ou d'organoïdes font partie des solutions innovantes à développer pour limiter les expérimentations animales en toxicologie par exemple.
Le potentiel de partenariat socio-économique de la biologie numérique dans les domaines de recherche d'INRAE est donc majeur. Pour DIGIT-BIO, la stratégie fera l'objet d'un travail d'élaboration de la part du comité stratégique, prenant en compte certains atouts de l'institut (e.g., Instituts Carnot Plant2Pro et France Futur Elevage, domaines d'innovation INRAE). Trois pistes de réflexion se dessinent : (i) renforcer des partenariats relativement matures par exemple dans les domaines de la sélection génétique, de la phénomique, et de la modélisation pour l'aide à la décision avec les partenaires semenciers et de l'AgTech (en lien par exemple avec l'Institut Convergence DigitAg); (ii) développer les partenariats avec les start-ups dans les domaines de la biologie numérique ; (iii) explorer les possibilité de partenariats dans les domaines de l'e-santé animale ou végétale.

6 – GOUVERNANCE ET PILOTAGE

La gouvernance de DIGIT BIO sera conforme au cadre proposé par le document « Principes et bonnes pratiques des MP INRAE » :

- Un binôme de direction associant un spécialiste en sciences formelles (H. Monod, CD MATHNUM) et une biologiste (C. Caranta), en charge de la stratégie scientifique et partenariale ainsi que du suivi du plan d'action du MP.
- Un.e chef.fe de projet sera en charge du suivi général du MP en appui de la direction, et contribuera à l'élaboration et la mise en œuvre des actions du MP (communication scientifique, organisation et suivi des appels à projets, analyse bibliométrique en lien avec l'IST, suivi des indicateurs, suivi du budget, etc…).

- Un comité de pilotage stratégique (CoStrat) constitué d'une quinzaine de scientifiques experts pour différents aspects de la biologie numérique, contribuera, aux côtés de la direction, aux orientations, aux animations et au suivi scientifique du MP. Le CoStrat associera également des scientifiques INRIA et CNRS. Il est attendu du CoStrat une contribution active et régulière à la vie du MP (e.g., analyse bibliométrique, élaboration

[11] Plant phenomics, from sensor to knowledge, https://doi.org/10.1016/j.cub.2017.05.055

des appels à projets et arbitrage, élaboration de notes/*position paper*, communication, construction/consolidation des partenariats nationaux et internationaux, etc…).

- Un comité scientifique international (*Scientific Advisory Board* – SAB) sera nommé, constitué de 4 à 6 experts internationaux reconnus dans le domaine de la biologie numérique. Il sera en charge d'émettre des avis et conseils sur les orientations, la stratégie et les avancées du MP, en particulier dans sa dimension internationale.

En termes de pilotage, les actions prévisionnelles au cours des 2 premières années sont les suivantes :

- Premier Trimestre 2021 : installation du CoStrat et élaboration conjointe des grandes étapes de la feuille de route 2021-2022 ;
- 2nd trimestre 2021 : séminaire de lancement et réseautage en vue d'initier la réflexion sur les trajectoires vers l'interdisciplinarité nécessaires pour répondre aux ambitions de DIGIT BIO ;
- A partir du 2nd trimestre 2021 : Mise en œuvre des leviers pour soutenir les trajectoires (projets exploratoires, réseaux, animations scientifiques, écoles chercheurs, équipes projets, missions internationales…), notamment via un appel à manifestation d'intérêt ;
- 1er trimestre 2022 : constitution et réunion du SAB.

7 – CRITERES DE SUCCES ET INDICATEURS ASSOCIES

Les actions qui seront soutenues par le MP DIGIT BIO concernent l'animation et le soutien aux communautés scientifiques et consortia interdisciplinaires (animations scientifiques, soutiens à des réseaux, constitution d'équipes projets, écoles chercheurs…), le développement de démarches (avec potentiellement une prise de risques) autour de projets exploratoires entre équipes et disciplines, la conception et la maturation d'un petit nombre de projets interdisciplinaires et emblématiques du MP et le soutien à des thèses sur des fronts de sciences interdisciplinaires.

Sur cette base, une attention particulière sera portée (i) à l'inscription de ces différentes actions au sein d'un parcours cohérent et construit vers l'interdisciplinarité s'inscrivant dans la durée, (ii) au potentiel « d'effet levier » des actions soutenues (e.g., actions permettant d'obtenir des résultats préliminaires pour ensuite déposer un projet à d'autres guichets, ou encore qui permettent d'intégrer un consortium européen ou international), et (iii) et à la mise en place d'animations dédiées à la « création collaborative numérique » (type hackathon) autour de grandes questions en biologie (e.g., Comment les organes savent quand s'arrêter de grandir?; Quel est le rôle de la stochasticité dans la signalisation en biologie?; Comment intégrer plusieurs feedbacks et leur contradictions?).

Dans l'objectif d'évaluer la trajectoire vers l'ambition et les objectifs affichés, mais aussi en tant qu'outil d'orientation, un certain nombre d'indicateurs (à consolider avec le comité d'orientation stratégique) seront suivis.

- Pour construire et animer une nouvelle communauté scientifique interdisciplinaire : nombre/thème des nouveaux réseaux soutenus, nombre/thème d'animations transversales et écoles chercheurs, nombre/thème d'animations collaboratives numériques,

nombre/typologie des collaborations interdisciplinaires, nombre de nouvelles équipes projet, nombre de scientifiques/disciplines/départements mobilisés.

- Pour augmenter la visibilité et la reconnaissance d'INRAE : Nombre de projets soutenus par le MP ou non (effet levier), productions scientifiques (publications, citations, review, …), thèses et devenir des doctorants, nombre/type d'actions de communication.
- Pour affirmer le positionnement d'INRAE sur la scène internationale : invitations dans des congrès internationaux, participation à des actions internationales (e.g. consortia, réseaux internationaux), nombre de soutien à des mobilités internationales de courte durée, nombre de visites dans instituts étrangers, productions scientifiques impliquant un partenaire étranger.