CID51
MODELISATION MATHEMATIQUE, INFORMATIQUE ET PHYSIQUE POUR LES SCIENCES DU VIVANT

Franck PICARD (président de section); Myriam FERRO (secrétaire scientifique); Alice CLEYNEN; Marie DOUMIC; Xavier DUCHEMIN; Guillaume FERTIN; Jonathan FILEE; Philippe JUIN; Romain KOSZUL; Rafael LABOISSIÈRE; Dominique LAVENIER; Thérèse MALLIAVIN; Antonio MONARI; Isabelle NONDIER; Françoise PEYRIN; Pierre POUGET; David VALLENET; Rufin VANRULLEN; Stéphane VEZIAN; Aleksandra WALCZAK.

Résumé

La CID 51 s'intéresse aux interactions entre les sciences de la modélisation (mathématique, informatique, physique) et les sciences du vivant. Ces interactions sont anciennes et se nourrissent des avancées de chaque discipline. Or la biologie, comme les sciences de la modélisation ont connu des avancées spectaculaires ces dernières années, technologiques et conceptuelles. Ces progrès rendent nécessaire la réflexion sur une nouvelle interdisciplinarité et sur l'exploration de défis scientifiques émergents aux interfaces qu'il semble crucial de relever. Ce rapport de conjoncture pour la période 2016-2021 est donc organisé en six axes qui n'ont pas vocation à être exhaustifs, mais qui ont permis aux membres de la CID51 de structurer une réflexion sur l'interdisciplinarité de demain. En préambule de notre rapport scientifique, nous proposons de dresser le premier portrait des chercheuses et des chercheurs rattachés à la CID51.

Introduction

En 2018 la revue Nature propose de recenser les 100 articles les plus cités toutes disciplines confondues : ClustalW (outil d'alignement multiple de séquences publié en 1994) obtient la 10ème place (40,289 citations), suivi par BLAST (12ème-14ème publiés en 1990 et 1997, 38,380 et 36,410 citations), l'algorithme du Neighbor Joining (20ème, publié en 1987), le bootstrap en phylogénie (41ème, publié en 1985), le false discovery rate (FDR, 59ème position publié en 1995). Ce classement illustre l'impact considérable des sciences de la modélisation aux avancées de la biologie de la fin du 20ème siècle. L'interdisciplinarité, dont le socle repose sur un dialogue constant entre disciplines scientifiques, s'est indéniablement renforcée au gré des avancées technologiques, et notamment de la massification de la production de données biologiques. Mais au-delà des aspects techniques, ce dialogue repose sur la capacité des sciences de la modélisation à faire émerger de nouveaux concepts permettant la meilleure compréhension des systèmes biologiques, et sur les défis posés aux sciences formelles par les sciences du vivant. Chaque discipline suivant sa propre évolution scientifique : les mathématiques, l'informatique, la physique et la chimie ont elles aussi connu des avancées importantes pouvant avoir des impacts en biologie, comme les grandes matrices aléatoires, et le transport optimal, l'optimisation combinatoire, et l'intelligence artificielle, la physique hors équilibre et les attracteurs continus, la chimie quantique, et les calculs d'énergie libre. Le rôle de la CID 51 est d'accompagner l'évolution de l'interdisciplinarité qui se transforme elle aussi, en identifiant, encourageant, et en promouvant les interactions qui potentiellement pourront faire émerger de nouveaux concepts aux interfaces entre les sciences biologiques et des sciences de la modélisation.

I. Les chercheurs et chercheuses de la CID51

Depuis 2009 les CID du CNRS suivent la carrière des agents qu'elles recrutent, et les étapes de la carrière de leurs lauréats hors la promotion. Après leur renouvellement en 2012, le SGCN a sollicité tous les agents recrutés par les CID entre 2009 et 2012 pour qu'ils choisissent une CID de rattachement. Certains ont choisi la CID51, d'autres n'ont pas répondu ou n'ont pas souhaité être rattachés à une CID. Depuis 2013, le Secrétariat Général du Comité National rattache systématiquement tout lauréat des concours à leur CID de recrutement, laquelle devient leur section secondaire. Fin 2018, la CID51 a lancé une enquête auprès des chercheuses et chercheurs rattachés et/ou recrutés par la CID51, pour dresser un portrait des membres de la CID (participation de 88%).

A. Portrait

En 2019, la CID 51 comportait 97 chercheuses et chercheurs, dont 60 sont CRCN (16 femmes), 30 DR2 (9 femmes), 7 DR1 (4 femmes) et 1 DRCE (0 femme). La répartition hommes/femmes est homogène selon les grades (~2/3 hommes, ~1/3 femmes). Les chercheurs rattachés à la CID 51 ont majoritairement (60%) été recrutés par la CID 51, ou par des sections associées : mathématiques 41 (6 chercheurs), neurosciences 25-26 (6), chimie 13-16-20 (5), physique 5 (3), et 75% des DR2 rattachés à la CID 51 ont été recrutés DR2 par la CID 51. Les DR1 ont quant à eux été promus en chimie (section 13-20, 3 promus), en biologie (section 21, 2), en mathématiques (section 41, 1) et en informatique (section 6, 1). Les membres de la CID 51 sont rattachés à titre principal aux sections d'informatique 6-7 (10 et 10 chercheurs respectivement), de mathématiques 41 (16) et de physique 2-5-11 (6-3-3). Le concours de recrutement concerne une moyenne de 61 candidats (CR) et 40 candidats (DR2) avec des taux de succès de 4 % et 8 % respectivement, ce qui en fait un des concours les plus sélectifs. L'âge moyen de recrutement (nombre d'années post thèse) a augmenté pour les CR, qui ont été recrutés en moyenne 5,5 années post thèse après 2015, alors qu'ils étaient recrutés en moyenne 4 ans post thèse entre 2005 et 2015. Les recrutements DR2 quant à eux interviennent plus tôt dans la carrière : 12,2 années post thèse pour les DR2 recrutés après 2015, contre 14,35 années sur la période 2005-2015, et 15 ans avant 2005. Lorsqu'on demande aux CR quelles sections ils envisagent pour le concours DR2, la majorité (90%) envisage de postuler en CID 51, et 43% envisagent de postuler dans une section en plus de la CID 51, génomique, biologie cellulaire et immunologie 21-22-27 (8-1-2), chimie 13-16-20 (3-2-2), écologie 29 (6), neurosciences 25-26 (2-4), informatique 6-7 (3-4), mathématiques 41 (3).

Enfin, les promotions DR1 ont quasiment toutes été obtenues après 2017 (6 promotions sur 7), et interviennent en moyenne 7,42 années après le recrutement DR2. Les CID ne faisant pas de promotion, les promotions DR1 dépendent essentiellement de la politique de promotion de l'interdisciplinarité dans les sections: parmi les 4 DR rattachés à la CID51 et dépendant d'une section de chimie à titre principal (13, 16, 20), 3 ont été promus DR1, alors que les sections de mathématiques et d'informatique n'ont promu que 1 DR1 sur 6 (mathématiques), et 1 sur 7 (informatique), aucun en physique (6 DR2), aucun en neuro (4 DR2) et 1 sur 6 en génomique (section 21-24).

Les membres de la CID 51 effectuent leur recherche dans des unités INSB (34%), INS2I (20%), INP (14%), INSMI (13%), INC (6%), INEE (6%), INSIS (3%), et deux chercheurs rattachés à la CID51 travaillent dans une unité INSERM. Deux tiers (63%) des sondés considèrent que leur laboratoire d'accueil est une structure de recherche interdisciplinaire. Les membres de la CID 51 accèdent aussi aux distinctions, avec 4 médailles de bronze, une d'argent (malgré l'absence de voie directe pour la désignation de médailles par les CID), et plusieurs prix de sociétés savantes.

B. Financements

Concernant les sources de financement sur projet sur les 5 dernières années, les membres de la CID 51 ont déposé un total de 127 projets ANR en tant que coordinateur, 164 en tant que responsable de tâche, et 91 en tant que collaborateurs, avec des taux de succès respectifs de 26%, 37% et 45%. On peut comparer ces chiffres au taux de succès de la nouvelle CES 45 de l'ANR (*Mathématique, Informatique, automatique, traitement du signal pour répondre aux défis de la biologie et de la santé*) qui était de 16,5% en 2018. Le taux

de succès des jeunes chercheurs (thèse postérieure à 2008) est quant à lui de 26% pour les coordinateurs. Ensuite, un tiers des chercheurs de la CID 51 a déjà eu accès à des financements nationaux hors ANR, qui proviennent pour un tiers des organismes de recherche biomédicale (FMR, INCA, Plan Cancer, INSERM, ARC), indiquant qu'une part non négligeable de l'activité des membres de la CID 51 est orientée vers les applications biomédicales. Pour ces appels, il ne nous est pas possible de déterminer si les projets concernaient des travaux méthodologiques ou applicatifs en collaboration avec des biologistes. Les financements CNRS font également partie des financements accessibles aux membres de la CID 51, le dispositif PEPS étant le plus répandu (34 projets). La moitié des chercheurs ont déjà eu accès à un financement relatif aux missions pour l'interdisciplinarité du CNRS, mais ces dispositifs sont critiqués pour le manque de lisibilité des modalités et des possibilités de dépenses, avec des calendriers souvent trop contraints. Leur attrait réside principalement dans la souplesse de l'appel d'offre. Est mentionnée également la multiplicité et donc le manque de lisibilité de l'ensemble des appels (PEPS, Osez l'interdisciplinarité, DEFI, PICS, Mission interdisciplinarité). Globalement, le fonctionnement de la mission pour l'interdisciplinarité semble mal compris par les chercheurs de la CID 51. Les structures universitaires financent également les chercheurs de la CID 51, puisque près de 60% ont déjà obtenu des financements provenant d'IDEX, de LABEX ou de chaires, ce qui indique que les financements locaux permettent de soutenir la recherche interdisciplinaire.

Les membres de la CID 51 demandent et obtiennent également des financements à l'ERC (starting : 3 sur 17 demandes, consolidator 2 sur 10, advanced 0 sur 1). Globalement, les financements ERC sont perçus comme étant souples et efficaces, même si trop compétitifs. D'autres instruments européens hors ERC financent également les membres de la CID 51, principalement des ETN (4 obtenus pour 8 demandes), ITN (3 obtenus pour 4 déposés), et des projets FET (2 obtenus pour 3 dépôts), mais le rôle respectif des membres de la CID51 n'est pas établi.

Seulement 40 % des chercheurs de la CID 51 considèrent qu'au cours des 5 dernières années, ils ont obtenu des financements suffisants pour que leurs recherches soient compétitives sur le plan international. L'enquête n'a pas permis d'établir le montant ni la répartition des financements. Le taux insuffisant de financement est rappelé, rejoignant ainsi l'ensemble de la communauté scientifique. La spécificité de l'évaluation des projets interdisciplinaires apparaît comme un point central pour garantir la crédibilité des décisions de financement. A ce sujet on peut noter que l'ANR a récemment mis en place la CES 45 dont les thématiques correspondent au cœur de métier de la CID 51.

Les membres de la CID 51 montrent un intérêt sur les financements interdisciplinaires s'appuyant sur des appels à projets succincts et ciblés, demandes qui permettent l'embauche de personnel, notamment pour les thèses interdisciplinaires. Un point central est le financement des personnels techniques sur les projets interdisciplinaires à moyen et long terme, car l'enquête montre que la majorité des chercheurs de la CID 51 estime que les financements actuels ne permettent pas de financer correctement les techniciens et ingénieurs. Enfin, les contraintes institutionnelles (entres instituts du CNRS par exemple), souvent difficiles à déchiffrer, sont vues comme un frein au dynamisme des projets interdisciplinaires.

C. Valorisation

Concernant la valorisation, environ 60 % des chercheurs de la CID51 n'ont jamais fait appel au CNRS pour la valorisation (brevet, logiciel etc.) ou la propriété intellectuelle de leurs travaux, et la moitié de ceux qui y ont fait appel sont satisfaits. L'enquête fait ressortir une demande d'amélioration de l'accompagnement pour la valorisation des activités spécifiques des membres de la CID 51, notamment pour les logiciels. Actuellement, les chercheurs de la CID 51 font souvent appel à d'autres structures (SATT) ou d'autres organismes (INSERM, INRA, INRIA). Est notée cependant une amélioration grâce à la mise en place de CNRS innovation, qui ne semble toutefois pas très connu de l'ensemble de la CID.

II. La biologie face aux défis des grandes masses de données

A. Des avancées technologiques majeures

La biologie moléculaire et cellulaire des dernières années est marquée par l'accessibilité remarquable des données à haut débit : les technologies de séquençage sont désormais utilisées en routine à bas coût, pour le séquençage de génomes entiers, la métagénomique, la détection de variants génétiques, la quantification de la transcription et de sa régulation. Les technologies récentes mettent l'accent sur des lectures plus longues, permettant de réduire les problèmes d'assemblage (au prix d'un nombre d'erreurs plus élevées), mais aussi l'identification de transcrits (notamment épissés). L'accessibilité croissante de ces technologies permet le développement de plateformes d'analyse dans les laboratoires et les établissements hospitaliers. De nouvelles attentes émergent car les praticiens ont besoin de standards de stockage et d'analyse pour s'inscrire dans une démarche de recherche reproductible à temps long. C'est un défi immense étant donné les contraintes importantes de confidentialité et de gestion des données (notamment en santé), dans un contexte où les technologies évoluent rapidement, et où les biologistes doivent se former pour appréhender les enjeux méthodologiques de l'exploitation de ces données.

Ces avancées techniques accompagnent deux sauts technologiques majeurs des années récentes : la possibilité de séquencer des molécules uniques et la possibilité d'accéder aux caractéristiques moléculaires de cellules uniques. Il s'agit désormais d'accéder à « l'identité » d'une population de cellules, sur la base des quantifications individuelles de leur génome, transcriptome, épigénome, et protéome. Ce développement de techniques de séquençage s'accompagne également de l'explosion de l'utilisation d'imagerie cellulaire, pour combiner les informations moléculaires (séquençage in-situ) et tissulaires (microscopie confocale à haute résolution). Les données d'imagerie sont désormais incontournables pour l'étude des processus physiopathologiques, avec de nouvelles techniques d'imagerie quantitative multi-échelles et multi-contrastes (la cryo-microscopie à résolution atomique aura certainement un impact considérable en biologie et bioinformatique structurale). Le développement de l'imagerie compressive est très prometteur, en intégrant la co-conception de l'instrumentation et de la chaîne de traitement des données au niveau même de la formation de l'image, ouvrant de nouvelles possibilités d'imagerie multi-spectrale ou vidéo temps réel.

B. Les enjeux des bases de données biologiques et biomédicales

La constitution et le maintien de bases de données consultables et exploitables est un enjeu qui accompagne les progrès technologiques, et qui doit être mis désormais en perspective des besoins considérables des techniques d'apprentissage automatique, dont les performances dépendent en grande partie des données d'entraînement correctement annotées. Les compétences en jeu concernent le matériel informatique stricto sensu ainsi que la recherche en informatique pour mettre au point des bases de connaissances robustes et consultables par tous. Face au déluge de séquences générées en génomique, ainsi qu'à la production massive d'images (par exemple en neurosciences), les communautés doivent s'adapter en utilisant à la fois les bases centralisées (dont le modèle est remis en question face à la charge matérielle et financière du maintien de ces banques), ainsi que de nombreuses bases plus locales et spécialisées, chacune reposant sur des formats standard, et s'inscrivant dans la logique de la science ouverte, qui devient une condition nécessaire à la publication.

Les défis considérables concernent l'hétérogénéité des données stockées, notamment en santé : séquences, images, données physiologiques, suivi temporel. Définir un modèle de développement informatique pérenne semble un défi majeur des prochaines années pour coordonner l'acquisition de ces données, leur stockage, leur annotation, leur accessibilité, leurs traitements, tout en respectant les normes qualité, qui peuvent être spécifiques, avec des contraintes

importantes en santé notamment concernant la traçabilité et la confidentialité. Ces enjeux s'inscrivent dans des axes de recherche informatique d'actualité relatifs aux grandes masses de données. L'ingénierie des connaissances, la fouille de données, le web sémantique, l'interconnexion de réseaux complexes sont au cœur des problématiques soulevées par l'exploitation de ces bases de données.

La sécurité des données et des réseaux, ainsi que la protection de la vie privée sont également des verrous à lever dans ce domaine. Le niveau de sûreté nécessaire pour faire du diagnostic n'est en effet pas comparable à celui rencontré habituellement en biologie prédictive. Ces exigences s'accompagnent d'aspects éthiques et de respect de la vie privée liés à l'exploitation bioinformatique de données individuelles à grande échelle (differential privacy). Il nous faut imaginer des environnements non centralisés, interopérables, qui puissent gérer cette hétérogénéité tout en assurant la confidentialité des patients. C'est l'une des clés pour exploiter de manière optimale la richesse de ces données.

Enfin, l'exploitation efficace de ces masses de données ne peut se passer du calcul haute performance (HPC) pour les traitements à haut débit, ce qui pose des défis considérables en termes d'infrastructures de calcul, de stockage de l'information, d'activité de service pour les laboratoires producteurs de données, ainsi que de production de logiciels d'analyse. La concurrence internationale est extrêmement sévère dans ce domaine, car des centres comme le Broad Institute ou l'EBI et le NCBI disposent d'une force de frappe considérable en termes d'ingénieurs et chercheurs en bioinformatique, ce qui leur permet de produire des méthodes pertinentes rapidement, et d'imposer ensuite leurs standards d'analyse.

C. Vers une nouvelle science des données biologiques et biomédicales

L'analyse de très grandes masses de données est un domaine de recherche extrêmement pointu et dynamique ces dernières années, grâce à la synergie entre différents domaines des mathématiques (statistique, optimisation, physique statistique, modélisation probabiliste ou déterministe) et de l'informatique (intelligence artificielle, fouille de données, algorithmique des séquences, représentation de connaissances). L'enjeu est le suivant : comment exploiter au mieux le potentiel de l'ensemble des données générées, comment les représenter pour mieux comprendre leur hétérogénéité, comment inférer les paramètres de modèles mathématiques ou physiques permettant de mieux comprendre le fonctionnement des systèmes biologiques, et comment intégrer les différents niveaux d'organisation biologique dans les modèles ? La représentation et la visualisation des données posent des défis considérables qui sont partagés par de nombreuses disciplines. Les visualisations linéaires multidimensionnelles classiques sont désormais supplantées par les méthodes non linéaires qui font appel à des notions complexes de géométrie, avec le développement de méthodes d'analyse topologique des données qui connaissent un essor remarquable et très prometteur. Ces données quantitatives nécessiteront ensuite des analyses statistiques et mathématiques pour inférer les réseaux ou les structures sous-jacentes en utilisant par exemple la théorie des graphes, la théorie du signal et la physique statistique. Afin de construire une vision intégrée des systèmes biologiques, il sera important de faire le lien entre les différentes échelles d'observation à l'aide d'approches de modélisation dynamique, telles que des équations algébriques, différentielles, aux dérivées partielles et probabilistes, ou encore des modèles d'agents (individus-centrés). L'étude des interactions, qu'elles soient moléculaires (réseaux de gènes) ou cellulaires (différenciation, signalisation), devra faire appel à des méthodologies innovantes qui incluent notamment les récents développements des méthodes d'inférence causale. Aussi, la disponibilité de données moléculaires sur des populations entières de cellules rend possible l'étude des états cellulaires en tant que continuums, faisant appel à de nouvelles approches fondées sur le transport optimal pour étudier les transformations des états de différenciation cellulaire. Le défi méthodologique est bien devant nous pour appréhender cette complexité jamais rencontrée, notamment grâce à la quantification de la stochasticité des phénomènes biologiques à grande échelle. Il ne pourra être relevé qu'en conjuguant les forces des différentes approches physiques, mathématiques et informatiques, pour faire

émerger une véritable multidisciplinarité autour des problématiques de biologie et de santé.

D. Apprentissage pour la biologie moléculaire et la santé

Parmi l'ensemble des approches mathématiques et informatiques mises en œuvre pour répondre au défi de l'analyse des grandes masses de données, les méthodes d'apprentissage automatique suscitent de nombreuses espérances. Le cadre de ces méthodes est d'apprendre une relation entre des variables de sorties (y) et des variables d'entrée (x), par un modèle de type $y=f(x)$, sans spécifier la forme de la fonction f à inférer. Ces méthodes apparaissent particulièrement intéressantes dans les cas où le fonctionnement du système biologique n'est pas connu précisément. Une spécificité des méthodes d'apprentissage automatique est leur objectif profondément prédictif : elles cherchent à prédire y en fonction de x, comme par exemple la prédiction de l'expression des gènes à partir de caractéristiques génétiques ou épigénétiques. Un des enjeux majeurs de l'apprentissage automatique est de sélectionner l'ensemble des variables (ou *features*) sur lesquelles reposent l'apprentissage d'une règle de prédiction performante, par exemple apprendre quels sont les gènes dont l'expression est prédictive de la survie d'un patient. Alors que beaucoup de développements ont été proposés dans les années 2000 avec le développement de la statistique en grande dimension, les réseaux de neurones ont connu un regain de popularité grâce au développement de nouvelles architectures de calcul. La puissance de ces approches est de proposer des méthodes computationnelles pour l'apprentissage de la fonction f en transformant les données d'entrée x en plusieurs étapes. Mais l'apprentissage automatique s'est surtout développé pour l'analyse d'images et de textes, dont le transfert des compétences à la biologie est un enjeu important. À ce titre, l'imagerie cellulaire et cérébrale bénéficie d'ores et déjà d'un corpus de méthodes extrêmement efficaces pour faire face à l'imagerie à haut débit. Une connexion directe avec la biologie moléculaire s'est opérée par le biais de l'analyse de séquence, car les techniques mises au point, comme les réseaux de convolution, pouvaient être transposées sans trop de difficulté. Un des défis reste cependant le développement d'architectures de réseaux qui soient dédiés aux problématiques biologiques, avec également l'utilisation d'architectures de calcul GPU et de cadres de développement (TensorFlow, PyTorch) non conventionnels pour les laboratoires de biologie. Les résultats de l'apprentissage automatique en biologie ces dernières années sont très prometteurs et soulèvent des défis théoriques majeurs. Un des principaux blocages concerne l'identification des variables ayant permis l'apprentissage du modèle, car l'objectif profondément prédictif des méthodes d'apprentissage ne permet pas toujours d'expliquer et d'interpréter les résultats biologiques. Coupler modélisation mathématique, inférence causale et apprentissage statistique pourrait permettre de dépasser ce blocage.

III. Biologie Intégrative et des systèmes

La biologie des systèmes émerge dans les années 2000, notamment grâce aux technologies permettant la quantification de phénomènes moléculaires et cellulaires à différentes échelles. La notion de complexité émerge de ces différentes phases de collecte de données, car le défi est d'intégrer ces différentes strates d'information pour mieux comprendre le fonctionnement des systèmes biologiques qui sont hétérogènes, dynamiques, variables, en évolution et en interaction constante avec un environnement biotique et abiotique. Les approches multidisciplinaires qui combinent mathématiques, informatique et physique ont alors un rôle central pour mieux comprendre la dynamique, l'évolution et le contrôle des systèmes vivants.

A. Les systèmes biologiques, des systèmes physiques complexes

La biologie des systèmes est profondément ancrée dans la démarche de modélisation physique, qui permet l'analyse de l'émergence des propriétés des systèmes à partir du comportement des éléments qui les constituent et des couplages entre leurs différents niveaux d'organisation. Comment les organismes vivants s'inscrivent-ils dans les lois de la physique ? Leur fonctionnement implique des processus uniques au vivant : ils se reproduisent et évoluent, ils perçoivent l'environnement en

transmettant des informations au monde extérieur, ils prennent des décisions leur permettant d'interagir avec leur environnement. La contribution unique de la physique est de définir et d'intégrer des contraintes et des limites aux mécanismes possibles d'évolution des systèmes vivants, telles que les limites imposées par le bruit moléculaire, les coûts énergétiques, la diffusion, les échelles de longueur et de temps. Les systèmes vivants mettent en œuvre des solutions fiables et souvent reproductibles, et ce à toutes les échelles, des récepteurs aux voies de signalisation et circuits génétiques régulant le phénotype et la différenciation cellulaire, puis aux populations évolutives et groupes d'animaux, en passant par les tissus et les réseaux de neurones. Un des objectifs de la physique biologique théorique multi-échelles est d'aller au-delà de la simple modélisation du comportement observé mais de trouver les lois et les règles qui permettent des théories prédictives et si possible unificatrices. Une direction novatrice est la formulation des lois phénoménologiques qui cherche à extraire les variables et interactions qui gouvernent les comportements observés au lieu de modéliser chaque interaction de signalisation.

Un des défis considérables de la biophysique moderne est d'extraire les mécanismes biologiques à partir de données toujours plus volumineuses et complexes, puis de les vérifier quantitativement au regard des données. Les interconnexions entre biophysique, statistique et apprentissage automatique sont donc extrêmement prometteuses mais constituent des défis méthodologiques majeurs (par exemple, l'interprétabilité des modèles, l'apprentissage des modèles dynamiques, la cohérence des modèles d'apprentissage avec les lois de la physique). La physique moderne du vivant utilise désormais des méthodes à l'origine développées en informatique, en théorie de l'information et en théorie de contrôle, en allant souvent plus loin dans l'interprétation pour faire un lien entre les résultats de ces approches et les mécanismes physiques. Ces potentialités se sont déjà révélées prometteuses pour l'étude de la chromatine, pour les attracteurs continus en neuroscience, l'étude des ligands immunologiques, la séparation de phase dans les tissus et la description du régime d'interférences de clones dans l'évolution. Actuellement, nous manquons de modèles bien établis en physique biologique théorique, ce qui en fait un domaine jeune et passionnant, où l'analyse statistique avancée des phénomènes observés à travers les échelles conduit à des modèles abstraits, dont les prédictions sont testées dans l'esprit de la physique.

IV. Recherche biomédicale et santé

La recherche biomédicale s'est récemment transformée grâce à l'accès facilité au séquençage et à l'imagerie à haut débit, le numérique devenant incontournable dans les systèmes de santé. La volonté de développer les techniques de e-santé et la mise en place accélérée des plateformes de séquençage au sein des hôpitaux constituent des atouts indiscutables, nécessitant cependant une réelle réflexion (notamment méthodologique) sur la mise en place de suivi longitudinal de cohortes, le stockage et l'analyse des données, et l'exploitation des résultats en clinique.

Des projets de recherche biomédicale d'envergure ont récemment émergé, comme le Plan Médecine Génomique 2025 (démarrage en 2016 avec le projet de financer 12 plateformes génomiques pour le diagnostic), ou le programme 3IA (lancé en 2019) avec sa composante santé. En imagerie, les infrastructures nationales France Life Imaging et France Bio Imaging ont été mises en place dans le cadre des Projets d'investissement d'Avenir dès 2013, avec des initiatives au niveau Européen également (European Institute for Biomedical Imaging Research). Au niveau international, de nombreux projets ou consortiums ont été lancés, certains avec participation de la France, comme le Brain Human Project, d'autres sans participation française, comme l'International Cancer Proteogenome Consortium. Tous ces projets partagent la caractéristique commune d'être interdisciplinaires et de mobiliser des expertises importantes en modélisation, analyse et gestion de données, ainsi qu'en intelligence artificielle. D'importants besoins en développements méthodologiques sont ainsi apparus afin, *in fine*, d'améliorer la compréhension de processus pathologiques mais également pour le bénéfice des patients (diagnostic, suivi thérapeutique) avec notamment l'émergence de la biologie des systèmes dans les hôpitaux.

A. Suivi de patients et médecine personnalisée

La modélisation de cohortes est un défi majeur de la recherche biomédicale translationnelle. Les études rétrospectives ont laissé la place aux études *on-line* intégrant les patients dès leur entrée dans l'hôpital. La quantité et l'hétérogénéité des données modernes de santé (séquences, signaux EEG, imagerie), ainsi que l'explosion des valeurs manquantes posent des difficultés dans la modélisation du suivi longitudinal de patients. L'apprentissage automatique commence à investir ces questions importantes, avec des résultats prometteurs, surtout dans les pathologies utilisant l'imagerie médicale ; toutefois, cette technique souffre drastiquement du peu de données d'apprentissage disponibles, et ne permettent que peu l'interprétation des résultats, question pourtant capitale pour comprendre les mécanismes des pathologies et en tirer les informations nécessaires à la prise en charge des patients. Il est donc important de lever ces enjeux en proposant des modèles robustes, flexibles et interprétables. La conception même des études à but d'intégration de données (moléculaires, phénotypiques, signaux ou images) reste une question ouverte, ce qui nécessite la prise en compte de la disponibilité des techniques et des patients, ainsi que des coûts des expériences. Leurs analyses, reposant sur des outils de normalisation des données, sont souvent adaptées de techniques en techniques, puis intégrées dans des logiciels clés-en-main, sans réelle remise en question de leurs performances. Des approches différentes se développent, basées par exemple sur la considération d'espaces de géométrie différente, ou sur des techniques de transport optimal.

Les avancées de biologie fondamentale amènent aussi à préciser les modèles physiologiques utilisés dans les domaines biomédicaux (poumon, sang, cœur). Les modèles biomécaniques de tissus vivants ont conduit par exemple à mieux simuler les croissances cancéreuses sur la base d'images de patients, à adapter les simulations de biofluides en s'appuyant sur une reconstruction individualisée des caractéristiques physiques du patient.

B. Des systèmes pathogènes stochastiques et dynamiques

Les interactions entre sciences biomédicales et modélisation ont récemment connu de nouveaux développements permettant de mieux intégrer expérimentation biologique, modélisation et applications médicales. Par exemple, l'étude de l'évolution du cancer a connu de nouvelles avancées grâce aux modèles mathématiques et physiques permettant de décrire et de prédire la croissance tumorale en interaction avec son environnement. Les approches de modélisation permettent désormais d'adopter une vision systémique de la tumeur en interaction avec le système immunitaire et le métabiome. Les modèles stochastiques ou à base d'équations aux dérivées partielles (au sens large) sont au cœur de la prédiction la dynamique de ces systèmes. En parallèle, les technologies d'imagerie étant de moins en moins invasives et plus précises, décrire les interactions cellulaires et identifier la répartition des contraintes et des forces dans ces systèmes dynamiques constituent un axe de recherche actif. Les théoriciens travaillent en étroite collaboration avec des expérimentateurs pour comprendre à la fois le rôle de la géométrie globale et locale, de la mécanique et l'utilisation de l'énergie dans le développement, la différenciation cellulaire, et la formation de tumeurs.

La modélisation dynamique, fondée sur des modèles d'équations différentielles ou des modèles stochastiques, est également au cœur des développements récents pour l'étude des dynamiques d'infections virales ou bactériennes. Physiciens et mathématiciens introduisent actuellement des idées issues de la théorie du contrôle et proposent des protocoles de vaccination plus efficaces et en cours de test. Dans tous ces cas, des notions avancées issues de la théorie des probabilités, de la statistique bayésienne et de la physique des événements rares sont utilisées pour aller au-delà des méthodes classiques de biologie computationnelle. Les formidables succès des modèles biophysiques, combinés au développement de nouvelles techniques d'inférence, ont montré que l'évolution globale de la grippe peut être prédite à des échelles de temps inférieures à un an. À plus grande échelle, les spécialistes des réseaux combinent différents types de données à grande échelle (contrôle des maladies, suivi des téléphones portables) pour améliorer les modèles épidémiologiques et pour aider à comprendre la propagation de maladies. Les tendances

récentes incluent des approches multi-échelles, dans lesquelles les données de séquençage sont mélangées avec des informations sociologiques. Cette recherche a des implications pratiques pour la santé publique, mais également pour les fondements mathématiques et physiques des systèmes en désordre avec des interactions.

V. Biodiversité, évolution et écologie

A. Génomique environnementale à haut débit

L'accessibilité croissante des données moléculaires à haut débit a provoqué une révolution dans l'étude de la diversité et de l'évolution des écosystèmes, avec des impacts majeurs en santé, climat, agriculture, et biodiversité par exemple. La génomique environnementale à haut débit a pour ambition d'identifier les acteurs de la biodiversité en matière d'espèces et de fonctions, en combinant données génomiques, génétiques, métaboliques, phénotypiques et données environnementales pour élucider les principes écologiques et évolutifs de base qui sous-tendent le fonctionnement d'écosystèmes complexes. Parmi les projets scientifiques à grande échelle en cours, nous pouvons citer l'exploration à l'échelle de la planète du plancton marin, qui est une composante clé du cycle du carbone et, également, l'étude du microbiote humain dont les interactions avec l'hôte sont un élément déterminant pour la santé. Cette révolution a été rendue possible grâce à la conception de structures de données spécialisées, pour indexer, interroger, fouiller, visualiser, structurer ou compresser ces masses de données. Ces avancées nécessitent une recherche algorithmique spécifique au domaine de la bioinformatique, qui trouve ses fondements en optimisation combinatoire, recherche opérationnelle, théorie des graphes ou algorithmique. Un enjeu particulier sera la conception de structures de données probabilistes, ainsi que de nouvelles heuristiques permettant le passage à l'échelle des méthodes d'indexation et de comparaisons globales des données de séquençage. Renforcer les liens entre informatique et statistique apparaît également crucial, notamment pour déterminer la significativité des variables d'intérêt détectées par des méthodes informatiques. Ces recherches s'accompagnent de validations théoriques et pratiques, en collaboration notamment avec des océanographes comme c'est le cas par exemple dans le cadre du projet Tara Océans. Enfin, la biophysique offre des perspectives prometteuses pour explorer les contraintes et les flux énergétiques dans les communautés et expliquer les liens entre le métabolisme, l'énergie, l'écologie et l'évolution. La physique (souvent hors équilibre) et les simulations numériques peuvent ainsi proposer des expériences d'évolution quantitative dirigées par les modèles, pour extraire des propriétés émergentes et prédire l'évolution et la robustesse du système face aux forçages environnementaux.

B. Modélisation des processus évolutifs et écologiques

La disponibilité de nouvelles technologies basées sur le séquençage, ainsi que des outils de surveillance en temps réel de communautés contrôlées, ont profondément transformé les approches éco-évolutives, en soulignant le besoin de nouvelles approches théoriques. Les lois fondamentales de l'évolution sont simples : les mutations génèrent des variations, tandis que la dérive, la recombinaison, la migration et la sélection génétiques modifient les fréquences des variants. Cependant, même dans des situations très simples, il est souvent étonnamment difficile de prédire comment ces forces agissent sur des millions d'individus afin de déterminer collectivement l'évolution d'une population. Physiciens, statisticiens et mathématiciens ont contribué de longue date en écologie-évolution (modèles de Lotka-Volterra, de populations structurées, de dérive génétique et plus récemment des apports de la théorie des matrices aléatoires). Les sciences formelles doivent désormais développer de nouveaux modèles intégrant des processus évolutifs comme la recombinaison, ainsi que des processus écologiques, pour étudier dans la dynamique d'adaptation et la structure des populations, la cohésion taxonomique et l'évolution dans des environnements fluctuants. En effet, le défi méthodologique consiste à intégrer un environnement intrinsèquement dynamique, tel que l'interaction éco-évolutive dans l'écologie microbienne, les interactions

système immunitaire-pathogène ou les tumeurs. Ces approches quantitatives permettront également d'étudier le contrôle et la prévisibilité des systèmes éco-évolutifs, qui sont des processus intrinsèquement aléatoires. L'augmentation spectaculaire de la quantité de données de séquences rend les comparaisons quantitatives avec les modèles prédictifs accessibles. Là encore, l'apprentissage automatique offre des perspectives prometteuses, avec la difficulté croissante d'identifier les bons niveaux de description des relations génotypes/phénotypes complexes.

La reconstitution de l'histoire évolutive des organismes et la phylogénie sont aussi fondamentalement transformées par l'accessibilité des génomes complets de nombreux organismes. Des progrès considérables ont été accomplis ces dernières années, alimentés par de nouveaux ensembles de données de séquençage et morphométriques et de données fonctionnelles. Les progrès méthodologiques accomplis combinent de nouvelles méthodes bayésiennes, souvent associées à des modèles statistiques d'évolution. Ces méthodes ont changé notre interprétation d'histoires évolutives, comme les phylogénies bactériennes, ou macro-évolutives. De nombreux travaux ont également été consacrés à la construction de modèles statistiques de corrélations dans les séquences existantes et à la séparation des contraintes évolutives et fonctionnelles. Cependant, les méthodes actuelles reposent souvent sur des phylogénies pré-assemblées, ce qui pose le défi de l'inférence phylogénétique prenant en compte la sélection, l'asymétrie de la production de descendants, les réarrangements à grande échelle, le transfert horizontal de gènes, les environnements variables et/ou structurés, y compris par exemple les taux élevés d'extinction-recolonisation de patches ou d'hôtes dans le cas des micro-parasites. Cependant, des travaux récents, combinant de grandes quantités de données et l'apprentissage automatique, proposent de nouvelles méthodes d'inférence, offrant de nouvelles perspectives au domaine.

VI. Biologie structurale et chimie du vivant

A. Intégration multi-échelle des phénomènes moléculaires

La biologie structurale a poursuivi ces dernières années le mouvement vers la biologie intégrative, qui tend à englober la description moléculaire des objets biologiques de l'échelle mésoscopique à l'échelle cellulaire. Cette tendance de fond implique, pour les aspects de modélisation moléculaire, le développement de modèles physiques permettant de prendre en compte simultanément plusieurs niveaux de détails dans la modélisation : tout-atome, gros-grains, ainsi que différents niveaux de modélisation quantique, qui aujourd'hui permettent non seulement de traiter la réactivité enzymatique mais aussi de décrire les phénomènes de photo-biologie. En parallèle, la démocratisation des architectures avec GPU donne accès à des temps de simulation et à des niveaux de théorie inimaginables il y a 5 ans.

La disponibilité de ces nouvelles ressources de calcul pose d'abord des défis informatiques d'efficacité pour les codes de modélisation moléculaire. Ces nouvelles ressources permettent l'accès à des durées de simulation de dynamiques moléculaires qui atteignent désormais de la micro-seconde jusqu'à la milliseconde (suivant la dimension des systèmes modélisés). La taille des systèmes étudiés peut maintenant aller jusqu'à un petit organisme (comme par exemple, un virion de 800 millions d'atomes). Par ailleurs, l'accroissement de la puissance de calcul induit un problème de taille des données, aussi rencontré dans les acquisitions de données cryo-EM (la taille d'un jeu des données étant de l'ordre du téraoctet). Des approches doivent être développées pour prendre en compte ces problèmes, en passant par des méthodes d'automatisation basées sur des analyses statistiques.

Le défi est de trouver le meilleur compromis entre la simulation de systèmes de plus en plus gros et la sophistication des modèles physiques de description des systèmes. Les deux aspects se complètent, car le paradigme basé sur des champs de force classiques utilisé jusqu'à présent nécessite d'être amélioré, par exemple pour prendre en compte la description fine des effets du pH et de la solvatation couplés avec les effets de la polarisation de l'environnement (champ de force polarisable). Ceci pourra notamment permettre un traitement plus réaliste des phénomènes de transfert de charge et

d'énergie (photosynthèse, respiration cellulaire), des modifications post-traductionnelles, ou alors la modélisation fine des échanges de protons pour les biomolécules avec la participation du solvant. Il s'agit là d'aspects qui sont cruciaux pour la description des processus cellulaires et physiologiques. Enfin, la simulation simultanée de modèles physiques avec différents niveaux d'approximation pose le problème du développement de théories physiques permettant de modéliser les interfaces entre ces niveaux, ainsi que l'interface entre les échelles moléculaire ou gros-grain et la description mésoscopique qui pourront se nourrir d'un dialogue plus poussé.

B. Approches multidisciplinaires pour la modélisation et la simulation moléculaires

Les verrous expérimentaux entre différents domaines qui relevaient des aspects moléculaires ou cellulaires sont progressivement levés. La multidisciplinarité induite par cette évolution pose le problème de la prise en compte dans les calculs reliés à la biologie structurale et moléculaire de données hétérogènes. Cette hétérogénéité provient de la diversité des méthodes utilisées aussi bien que de la complexité des phénomènes moléculaires sous-jacents aux processus physiologiques et cellulaires. Le désordre conformationnel d'un nombre de plus en plus grand de régions des protéines étudiées, ou l'hétérogénéité conformationnelle de la chromatine responsable en partie des difficultés d'analyse statistique des données de conformation en sont un exemple emblématique. La prise en compte de l'hétérogénéité des données nécessite le développement d'approches statistiques permettant d'appréhender cette diversité des données et des technologies. Dans ce sens, les approches fondées sur l'apprentissage automatique semblent constituer une piste prometteuse, et leur généralisation à l'analyse des trajectoires de dynamique moléculaire serait à envisager. L'introduction des approches bayésiennes pour l'ajustement de structures aux données expérimentales de biologie structurale sera certainement poursuivie et étendue à de nouveaux domaines, comme la tomographie aux rayons X ou bien la modélisation *ex novo* de structures de biomolécules ainsi qu'aux méthodes hybrides. De plus, un nouveau champ d'action pour les approches bayésiennes est représenté par la prise en compte de la dynamique interne des objets moléculaires.

Quant aux interactions entre macromolécules biologiques et ligands, la croissance de la puissance de calcul combinée à l'amélioration des méthodes d'échantillonnage permettent l'intégration des méthodes de scores des poses d'amarrage (docking) et des outils de modélisation moléculaire basés sur l'échantillonnage augmenté. De plus, les développements poussés des méthodes d'énergie libre ont pour but de prendre en compte la complexité de la définition des variables collectives dans les systèmes de complexité croissante, pour intégrer des stratégies basées sur les forces de biais et la méta-dynamique. Dans ce sens, la géométrie différentielle stochastique (provenant des mathématiques fondamentales) fournit des outils précieux pour la compréhension et la conception des méthodes d'échantillonnage augmenté.

Le développement de méthodes statistique de scores « ligand-based » pourra se baser sur les jeux d'apprentissage formés de mesures réalisées dans des conditions hétérogènes, afin de proposer des outils robustes. Une piste prometteuse pour améliorer les descripteurs des ligands est d'utiliser les avancées des méthodes de chimie quantique.

Enfin, il est absolument indispensable que la puissance de calcul disponible dans les centres de calcul nationaux ainsi que dans les laboratoires rattrape la progression actuellement observée dans les autres pays industrialisés, afin de ne pas faire subir un déclassement à l'ensemble de la communauté des bio-informaticiens en France. Dans ce cadre, si la proposition de la création de « data centers », considérée comme une priorité gouvernementale, est à saluer dans la mesure où elle peut pousser à l'augmentation des ressources de calcul, elle devra aussi être suffisamment flexible pour permettre une utilisation et exploitation optimales aux acteurs de la modélisation moléculaire et cellulaire. En sens complémentaire, les équipes interdisciplinaires dédiées à la modélisation moléculaire doivent pouvoir intégrer pleinement les opportunités offertes par la priorité stratégique donnée à l'intelligence artificielle.

VII. Neurosciences computationnelles

L'un des grands défis que doivent encore relever les neurosciences consiste à générer une compréhension non seulement quantitative, mais également fonctionnelle, voire algorithmique, des comportements émergents du système nerveux, de la sensation à l'action, en passant par la décision ou la mémoire. Parmi les caractéristiques encore mal comprises figurent notre capacité à percevoir, interpréter, apprendre et prédire le monde qui nous entoure, à créer des souvenirs qui peuvent durer toute une vie et à prendre des décisions qui nous rapprochent de la réalisation de nos objectifs. La cognition et le comportement sont la propriété émergente de l'activité de réseaux contenant des milliards de neurones, alors que nous commençons tout juste à comprendre les propriétés de calcul collectif de centaines de neurones, grâce à des techniques expérimentales de pointe associées à des approches informatiques et théoriques qui rassemblent des méthodes issues des statistiques, mathématiques, physique, imagerie et informatique. Ces problèmes doivent être étudiés à plusieurs échelles - des cellules et circuits simples jusqu'à la cognition et au comportement.

A. Multiples échelles spatiales et temporelles

L'un des défis principaux pour les neurosciences computationnelles est celui du passage à l'échelle, depuis le monde microscopique des molécules, neurotransmetteurs, dendrites et synapses vers le monde macroscopique des populations de neurones, des aires cérébrales ou des cerveaux et organismes entiers. Des modèles mathématiques du comportement d'un ou plusieurs neurones connecté(s) existent déjà, mais ne peuvent pas rendre compte de fonctions cognitives complexes. Les limites de champs moyens ont commencé à permettre de modéliser les assemblées de neurones sur la base du comportement physiologique d'un neurone. L'enjeu est de pouvoir extrapoler leur comportement au niveau d'une population, voire de plusieurs populations neuronales inter-connectées. De nouveaux outils mathématiques sont donc nécessaires, s'inspirant des modèles stochastiques ou des approximations de champ moyen. Ces outils méthodologiques sont d'autant plus nécessaires, et d'autant plus sophistiqués, que les techniques expérimentales d'où provient l'information biologique accroissent leur résolution et atteignent de très hauts débits (par exemple, enregistrements multi-électrodes, imagerie bi-photonique).

La question du passage à l'échelle concerne également les aspects temporels, avec à la fois des mécanismes de signalisation rapide (de l'ordre de la milliseconde) et des réponses électrophysiologiques plus lentes (plusieurs dixièmes de seconde), jusqu'à des fonctions cognitives s'exprimant sur une ou plusieurs secondes, et des processus de mémorisation à plus ou moins long terme, pouvant durer toute une vie. Le cerveau étant un système adaptatif par excellence, ces dynamiques temporelles et leur évolution sont souvent anticipées ou prédites au sein du système lui-même, donnant lieu à une représentation dynamique de l'environnement interne comme externe. Quels sont les algorithmes de calcul de la réponse et de son adaptation, quel est le traitement du signal dans ces conditions changeantes, souvent aléatoires ? A la plus large comme à la plus petite échelle spatiale et temporelle, des théories efficaces, basées sur des approches de physique statistique et des modèles mathématiques probabilistes (par exemple les approches bayésiennes, théorie d'information, modèles de dynamique stochastique non-linéaires, théorie du contrôle, inférence statistique), permettent de relier l'échelle fonctionnelle à des résultats vérifiables expérimentalement. La physique a également apporté des idées sur le codage de la mémoire dans des attracteurs tant discrets que continus. Les dernières années ont montré pour la première fois, avec l'aide de la théorie pour diriger et analyser les expériences, la preuve expérimentale d'attracteurs continus. Une exploration plus poussée dans d'autres domaines, ainsi que la compréhension de leur codage et de leurs conséquences, ouvrent de nouveaux horizons aux systèmes dynamiques en neurosciences.

B. Des neurosciences computationnelles à l'intelligence artificielle

Notre capacité de modélisation du système nerveux croît en proportion directe avec les progrès en informatique, que ce soit au niveau

matériel (e.g. processeurs graphiques), logiciel ou algorithmique. L'essor (ou le renouveau récent) de l'Intelligence Artificielle (IA) est à la fois moteur pour la modélisation en biologie, et l'un des premiers domaines d'application des découvertes en neurosciences computationnelles. L'apprentissage automatique offre des méthodes de plus en plus puissantes pour capturer, reproduire et modéliser des observations expérimentales de plus en plus détaillées (e.g. imagerie microscopique, électrophysiologie, imagerie cérébrale fonctionnelle). D'autre part, certains modèles d'intelligence artificielle (notamment ceux basés sur les réseaux de neurones profonds) peuvent être compris et analysés en tant que nouveaux systèmes "intelligents", et donner lieu à de nouvelles théories du traitement de l'information dans le cerveau, qu'il conviendra ensuite de valider expérimentalement. Enfin, les progrès théoriques réalisés dans l'étude du cerveau peuvent servir en retour à alimenter la recherche en IA : c'est ainsi que l'inspiration des neurosciences a permis d'établir les premiers modèles de réseaux de neurones il y a près de 50 ans, et ouvert la voie au "deep learning" actuel. De la même manière, les neurosciences computationnelles d'aujourd'hui doivent permettre l'émergence de nouvelles architectures de réseaux de neurones, de nouvelles méthodes de codage neuronal ou d'apprentissage automatique pour l'IA de demain. Un défi important pour les années à venir serait de pouvoir tirer parti des logiciels de différentiation automatique issus de l'IA (comme Tensorflow et PyTorch) pour les approches de neurosciences computationnelles.

C. Imagerie cérébrale et traitement du signal

Les progrès des méthodes expérimentales en anatomie ou électrophysiologie (nouvelles imageries microscopiques, imagerie multimodale, multi-échelle sur des grands volumes, définition de nouveaux contrastes) s'accompagnent de besoins toujours croissants en traitements du signal, traitements d'images, segmentation et reconstruction d'images, y compris pour la résolution de problèmes inverses (retrouver les causes d'un signal multi-dimensionnel, en reconstruire les sources, etc.). Dans le domaine de l'imagerie cérébrale, la notion de "connectome" fait référence à une description multi-échelles des populations neuronales et de leurs interactions, souvent manifestées par des connexions ou faisceaux de connexions axonales. L'analyse de ce nouveau genre de données massives requiert des techniques avancées pour le traitement de signal sur graphes. De manière générale, diverses méthodes statistiques (par exemple, réduction de dimension, modèles parcimonieux, techniques d'inférence Bayésienne) sont nécessaires pour répondre aux enjeux du "big data". Là aussi, les méthodes d'apprentissage automatique en général et de Deep Learning en particulier apportent de nouvelles solutions plus appropriées aux données massives. Un exemple est la parution récente de diverses méthodes de segmentation automatique de neurones, axones ou dendrites en imagerie calcique et biphotonique basées sur des réseaux convolutionnels profonds. Ces problématiques posent également la question de la constitution, de la gestion et du partage de grandes bases de données d'imagerie, déjà abordée précédemment.

D. Comportements et intelligences collectives

Enfin, le comportement peut être étudié sans passer à l'échelle du neurone (voire même pour des systèmes non-neuronaux, tels que des bactéries). On peut prendre pour exemple les mouvements collectifs de bancs de poissons ou de volées d'oiseaux, ainsi que l'intelligence distribuée de certains insectes "sociaux", telles les fourmis. Ces systèmes comportementaux ont été utilisés pour proposer des algorithmes de recherche et explorer le rôle des a priori codés, de la mémoire et de la détection. Les modèles récents de comportement collectif, fondés sur des données expérimentales, nous ont aidé à comprendre le rôle des interactions dans les systèmes dynamiques et à quantifier le mouvement complexe. Pour aller plus loin, il conviendra de combiner des modèles d'environnement (e.g. environnements turbulents), des expériences quantitatives ainsi que des théories sur l'orientation et la prévision dans les environnements stochastiques (rôle de l'apprentissage, rôle de la perception).

VIII. Recommandations

A. Ingénieurs, chercheurs, plateformes, structures nationales

La recherche méthodologique en modélisation pour la biologie se nourrit de l'interface entre les acteurs des sciences du vivant et les méthodologistes. Les plateformes de bioinformatique regroupées au sein de l'Institut Français de Bioinformatique (IFB) ont un rôle stratégique car elles assurent cette activité translationnelle. Elles permettent le transfert de l'expertise acquise sur la manipulation et l'annotation des données biologiques (étape cruciale pour toute méthode d'apprentissage notamment). L'activité du chercheur interdisciplinaire doit s'appuyer sur le support d'ingénieurs et techniciens en informatique et analyse des données. Ces besoins concernent la maintenance et le fonctionnement d'infrastructures informatiques (centres de calcul, serveurs, bases de données) mais également un travail d'ingénierie logicielle (et de maintenance à long terme), afin de rendre disponibles et accessibles les développements méthodologiques réalisés par les chercheurs (interfaces homme-machine, GUI), dans un contexte où les FAIR data (Findable, Accessible, Interoperable, Reusable) doivent devenir un standard. Par conséquent les recrutements d'ingénieurs et techniciens doivent se faire en cohérence avec ceux des chercheurs interdisciplinaires.

En plus de structures dédiées aux interfaces biologie/modélisation, il est absolument indispensable que la puissance de calcul disponible dans les centres de calcul nationaux ainsi que dans les laboratoires rattrape la progression actuellement observée dans les autres pays industrialisés. Dans ce cadre si la proposition de la création des « data centers », considérée comme une priorité gouvernementale, est à saluer dans la mesure où elle peut pousser à l'augmentation des ressources de calcul, elle devra aussi être suffisamment flexible pour permettre une utilisation et exploitation optimale aux acteurs de la modélisation moléculaire et cellulaire.

A ce titre les infrastructures nationales doivent être soutenues (TIMES, GENCI, CC-IN2P3, France Grilles, IFB), et les efforts nationaux doivent être coordonnés, notamment pour que les chercheurs français puissent mieux se positionner au sein de projets internationaux. Par ailleurs, les serveurs et centres de calcul induisent un coût énergétique/environnemental qui croît avec la masse des données. Il serait donc intéressant que les utilisateurs de ces structures aux interfaces avec la biologie entament une réflexion sur ces aspects.

Aussi, le CNRS doit renforcer les outils de valorisation et de soutien juridique aux chercheurs impliqués dans des recherches interdisciplinaires, notamment translationnelles. Ce soutien doit aussi s'accompagner d'une réflexion concernant l'impact des recherches et les responsabilités éthiques et morales des chercheuses et chercheurs aux interfaces biologie-modélisation, notamment en lien avec l'importance croissante des méthodes d'intelligence artificielle en biologie.

Enfin, l'animation scientifique de ces communautés s'articule autour de Groupements de Recherche interdisciplinaires qu'il est nécessaire de continuer à soutenir (comme les GDR BIM, Madics, IRN PAN, CellTiss, IAEM, IA, Médyna, MAMOVI).

B. Rassembler l'interdisciplinarité dans une unité de lieu

Le paysage français inclut quelques équipes de recherches et des plateformes interconnectées, mais la concurrence internationale est extrême dans ce domaine : plusieurs pays (USA, UK, Pays Bas) ont adopté de longue date une stratégie de structures intégrées rassemblant les différents acteurs de la recherche interdisciplinaires en biologie comme en santé. Promouvoir la constitution d'équipes interdisciplinaires (dans les laboratoires de biologie ou dans les laboratoires de mathématiques, informatique ou de physique) semble crucial pour développer une recherche interdisciplinaire ayant une visibilité internationale. Mais il faut trouver un équilibre entre quelques équipes pointues à effectif limité, et des centres de recherche dédiés à l'interdisciplinarité, qui sur le long terme ont la masse critique suffisante pour insuffler une vraie dynamique de recherche aux interfaces. La structuration du CNRS en instituts dont les chercheurs dépendent exclusivement apparaît comme une difficulté à la constitution

d'équipes de recherches interdisciplinaires. Une réflexion est à mener pour proposer des structures plus souples, au sein desquelles les chercheurs pourraient être recrutés sans contrainte d'appartenance à tel ou tel institut. Cette souplesse doit aussi concerner la formation doctorale en permettant aux étudiants de développer leur projet au sein de structures transverses.

C. Renforcer la CID51 en tant que pivot de la multidisciplinarité au CNRS

La CID51 est un outil stratégique pour le CNRS, car elle permet de recruter et d'évaluer des chercheurs de haut niveau aux interfaces. La constitution de jurys dédiés, composés à la fois de méthodologistes et de biologistes, doit absolument être maintenue car l'interdisciplinarité nécessite une inspection et une évaluation spécifiques des dossiers. Cela permet au CNRS de recruter des profils atypiques, d'un excellent niveau, qui n'auraient pas été sélectionné par une section disciplinaire (par exemple des candidats ayant changé de thématiques dans leur parcours, ou dont le projet aux interfaces aura plus de difficulté à être évalué par une section disciplinaire). Un des objectifs de la création de la première CID biologie-modélisation était de constituer une communauté de chercheuses et chercheurs aux interfaces. L'initiative, conçue comme étant ponctuelle et limitée dans le temps au départ, s'est avérée être un formidable outil de développement de l'interdisciplinarité au CNRS. Une véritable communauté de chercheurs est constituée, ce qui s'illustre par l'abondance des GDR consacrés à la modélisation (au sens large) en biologie et par la constitution d'équipes de recherche reconnues internationalement sur ce thème, dans différents instituts du CNRS.

Pourquoi maintenir une CID dans ce contexte ? Au vu de l'évolution de la biologie en tant que science de plus en plus quantitative, et de l'évolution des sciences de la modélisation d'autre part, la CID 51 doit être maintenue et renforcée pour accompagner les transformations des interactions biologie-modélisation. En effet, à l'inconnu inhérent aux découvertes disciplinaires s'ajoute l'inattendu des potentielles interactions entre disciplines. Nous pensons que cette véritable interdisciplinarité nécessite des structures particulières. Le statut de CID permet de rassembler plusieurs instituts autour d'une même politique scientifique de recrutements, ce qui est unique et précieux. Ce statut pourrait être réformé en termes de fonctionnement et de prérogatives. Concernant le fonctionnement, les modalités de recrutement des membres élus des CID apparaissent inadaptées alors que les mandats durent 5 ans : les membres des sections disciplinaires siégeant en CID participent à deux concours, ce qui représente un investissement déraisonnable sur une période aussi longue, ayant pour conséquence une difficulté à trouver des volontaires (et donc des experts). Concernant les prérogatives, la spécificité du recrutement et de l'évaluation par les CID doit pouvoir s'étendre à la promotion des chercheurs hors-classe et des directeurs de recherche 1ère classe. En effet, l'intérêt pour l'interdisciplinarité peut se manifester tout au long de la carrière, et il nous semble essentiel de promouvoir des DR1 ayant un profil interdisciplinaire, et qui auraient eux aussi opéré des changements thématiques ou acquis des compétences fortes dans un domaine autre que celui de leur expertise initiale. Enfin, pour promouvoir l'interdisciplinarité au CNRS, il semblerait judicieux que la CID 51 puisse proposer des médailles de bronze et des primes (PEDR) au titre de la recherche interdisciplinaire.