RAPPORT DE CONJONCTURE

DU COMITÉ NATIONAL DE LA RECHERCHE SCIENTIFIQUE

ÉDITION 2014

Extrait

CNRS ÉDITIONS

15, rue Malebranche – 75005 Paris

ISBN : 978-2-271-08746-1

CID 51

MODÉLISATION, ET ANALYSE DES DONNÉES ET DES SYSTÈMES BIOLOGIQUES : APPROCHES INFORMATIQUES, MATHÉMATIQUES ET PHYSIQUE

Extrait de la déclaration adoptée par le Comité national de la recherche scientifique réuni en session plénière extraordinaire le 11 juin 2014

La recherche est indispensable au développement des connaissances, au dynamisme économique ainsi qu'à l'entretien de l'esprit critique et démocratique. La pérennité des emplois scientifiques est indispensable à la liberté et la fécondité de la recherche. Le Comité national de la recherche scientifique rassemble tous les personnels de la recherche publique (chercheurs, enseignants-chercheurs, ingénieurs et techniciens). Ses membres, réunis en session plénière extraordinaire, demandent de toute urgence un plan pluriannuel ambitieux pour l'emploi scientifique. Ils affirment que la réduction continue de l'emploi scientifique est le résultat de choix politiques et non une conséquence de la conjoncture économique.

L'emploi scientifique est l'investissement d'avenir par excellence

Conserver en l'état le budget de l'enseignement supérieur et de la recherche revient à prolonger son déclin. Stabiliser les effectifs ne suffirait pas non plus à redynamiser la recherche : il faut envoyer un signe fort aux jeunes qui intègrent aujourd'hui l'enseignement supérieur en leur donnant les moyens et l'envie de faire de la recherche. On ne peut pas sacrifier les milliers de jeunes sans statut qui font la recherche d'aujourd'hui. Il faut de toute urgence résorber la précarité. Cela suppose la création, sur plusieurs années, de plusieurs milliers de postes supplémentaires dans le service public ainsi qu'une vraie politique d'incitation à l'emploi des docteurs dans le secteur privé, notamment industriel.

Composition de la commission interdisciplinaire – CID

Eduardo ROCHA (président de la CID) ; Michael BLUM (secrétaire scientifique) ; Luís ALMEIDA ; Pierre-Olivier AMBLARD ; Guillaume BESLON ; Isabelle CALLEBAUT ; Peggy CENAC-GUESDON ; Thibault COLLIN ; Henri COULAUD ; Florence D'ALCHÉ-BUC ; Alexandre DE BREVERN ; Audrey DUSSUTOUR ; Olivier GIPOULOUX ; Hinrich GRONEMEYER ; Daniel KAHN ; Arnaud LEJEUNE ; Sylvie MAZAN ; Jacques MILLET ; Hélène MORLON ; Yann PONTY ; Ovidiu RADULESCU.

Résumé

La biologie dépend de plus en plus de technologies capables de produire de vastes quantités de données. En conséquence, l'utilisation de méthodes sophistiquées de modélisation des connaissances et d'analyse des données est devenue essentielle pour répondre à des questionnements de plus en plus complexes sur le fonctionnement du vivant. Cela a amené au développement rapide de l'interface entre la biologie, les mathématiques, l'informatique, la physique et la chimie. Si la biologie tire de ces interactions une capacité accrue d'analyse de données et de synthèse des connaissances, les autres disciplines bénéficient en retour de nouveaux questionnements et d'exemples pour des nouvelles directions de recherche. Cette section détaille le niveau actuel de connaissances à l'interface entre la biologie et les autres disciplines. Elle identifie enfin certains obstacles au développement de l'interdisciplinarité en biologie et propose un ensemble de recommandations.

Introduction

La biologie est souvent présentée comme une discipline expérimentale, par opposition à d'autres disciplines disposant de solides fondations mathématiques comme la physique. Pourtant la modélisation a une longue tradition en biologie, comme l'attestent le modèle de la structure de l'ADN, les modèles d'activité électrophysiologique du neurone, ou le modèle d'allostérie en biochimie du métabolisme et de la signalisation. De même, l'anatomie comparée, le dogme central de la biologie moléculaire ou les formalisations mathématiques de la théorie de l'évolution ont eu un rôle majeur dans la structuration des connaissances en biologie. Plus récemment, la modélisation statistique du génome a contribué au succès des logiciels de bioinformatique. Malgré ces quelques exemples paradigmatiques, de par la complexité du sujet (le « vivant »), les modèles développés en biologie sont souvent restés limités à des schémas graphiques, éventuellement complétés par l'analyse statistique de données expérimentales. De fait, la biologie a été un des principaux moteurs du développement de la statistique. Réciproquement, les modèles biologiques, bien qu'ils soient longtemps restés assez élémentaires, utilisaient déjà des concepts et/ou des outils dérivés d'autres disciplines.

Depuis une vingtaine d'années, les représentations schématiques et phénoménologiques classiquement utilisées en biologie ont fait place à des développements beaucoup plus poussés impliquant les mathématiques, l'informatique, la physique ou la chimie, donnant ainsi naissance à une nouvelle communauté de recherche, positionnée à l'interface entre les sciences du vivant et une ou plusieurs de ces disciplines. Le champ thématique de la CID51 est justement situé à cette interface : les membres de la CID51 sont des chercheurs revendiquant une certaine interdisciplinarité et qui cherchent à apporter des réponses à des problèmes importants de biologie en développant des outils ou en appliquant des concepts issus d'autres disciplines.

Certains domaines de la biologie, comme la génétique des populations ou la biologie structurale, ont depuis toujours eu recours à des outils mathématiques ou informatiques. D'autres n'en ont eu la nécessité que depuis quelques années en raison du développement de méthodes d'acquisition de données à grande échelle. Comme le volume des données disponibles en biologie augmente rapidement, la bioinformatique et l'analyse des données deviennent essentielles pour extraire l'information pertinente des résultats expérimentaux.

Parallèlement, la modélisation s'est imposée comme un fantastique outil pour déduire logiquement des conclusions à partir d'un ensemble d'hypothèses. Ainsi les modèles peuvent servir à proposer des hypothèses – ou au contraire à les invalider, à prédire la réaction d'un système biologique à une perturbation ou à structurer la connaissance pour comprendre l'importance relative d'un ensemble de variables dans un processus biologique. La modélisation permet

ainsi une meilleure compréhension du vivant mais aussi une réponse beaucoup plus rapide à des enjeux sociétaux majeurs tels que l'épidémiologie des pathogènes émergents, la découverte de nouvelles approches pour le diagnostic et le traitement des maladies humaines ou le développement de la biologie de synthèse à visée industrielle.

Ce chapitre est divisé en trois parties. Les premières sections (I à V) présentent les domaines de la biologie qui ont un recours intensif aux techniques computationnelles et à la modélisation. Ensuite (sections VI à VIII), nous présentons spécifiquement les interfaces de la biologie avec la physique, les mathématiques et les sciences de l'information. Enfin, dans les sections IX à XI, nous concluons avec une description de la communauté, de ses atouts mais aussi des principales difficultés qui freinent le développement de l'interdisciplinarité et de la modélisation en biologie.

I. Génomique

La génomique constitue une rupture importante avec la génétique classique parce qu'elle permet d'étudier l'ensemble de l'information du génome dans un contexte intégratif. Depuis quelques années, de nouvelles technologies de séquençage ont ouvert la génomique à pratiquement tous les domaines de la biologie. Ainsi, le re-séquençage de génomes permet le développement de la génomique associative (recherche d'allèles associés à un phénotype) et des approches phylogénomiques en épidémiologie moléculaire. De nouvelles techniques permettent l'étude des interactions protéine-ADN, de l'expression des gènes, ou des variants fonctionnels. La génomique s'impose aussi en écologie où les approches de métagénomique révolutionnent l'écologie microbienne et le séquençage a un énorme impact sur la compréhension de la génétique et de l'histoire des populations naturelles. Les grands projets de génomique fournissent également des données pouvant être intégrées à différents niveaux dans des approches de biologie structurale. Les techniques de séquençage à haut débit génèrent des données qui permettent l'étude à basse résolution de la structure des chromosomes et l'analyse de l'impact de mutations sur la structure et fonction des protéines. Enfin, d'autres techniques permettent d'étudier les interactions ARN/protéines, ou encore de caractériser la structure secondaire des ARN. Toutes ces applications ont connu des développements importants durant les quatre dernières années et ont contribué à la croissance exponentielle et la diversification de l'information génomique contenue dans les bases de données. La France s'est initialement bien positionnée dans le domaine de la génomique. Même si elle a pris du retard dans les infrastructures de séquençage et dans la formation et le recrutement de bioinformaticiens, la France reste le quatrième contributeur mondial aux revues les plus citées en génomique.

La dissémination rapide de nouvelles technologies de séquençage plus performantes, moins chères, permettant des lectures plus longues et/ou à partir d'échantillons plus petits, pose des problèmes majeurs. Certaines de ces nouvelles méthodes permettent le séquençage de cellules uniques, mais produisent des séquences de faible qualité qui posent des défis algorithmiques pour l'assemblage et l'analyse. Elles apportent cependant un avantage décisif en permettant d'étudier la variabilité génomique au sein d'une population. L'individualisation de l'étude des génomes favorisera l'utilisation de techniques de génétique des populations et de physique statistique pour étudier, dans des populations hétérogènes, l'effet conjoint de l'expression génétique et de la dynamique des génomes sur le phénotype.

Les progrès de la génomique ont contribué à l'essor de nouvelles disciplines comme la génomique associative. Celle-ci utilise des approches statistiques pour identifier les régions du génome expliquant la variation de traits phénotypiques. Elle est donc d'un intérêt médical majeur. Cependant, en dépit du nombre impressionnant d'études conduites

dans les dernières années, le pouvoir explicatif et prédictif des variants détectés est souvent resté assez faible. Les efforts récents dans ce domaine concernent donc l'intégration de données de différentes natures : génomique, épigénomique, structurales et fonctionnelles, avec comme objectif principal de mieux comprendre le déterminisme biologique des pathologies. Ces nouveaux défis de la génétique associative sont de beaux exemples de recherches interdisciplinaires puisqu'ils font appel à des méthodes de statistiques de pointe ainsi qu'à des technologies post-génomiques nouvelles. De plus, ils répondent à des enjeux sociétaux importants en santé, typiquement en cancérologie, mais aussi sur les maladies neurodégénératives ou les maladies rares.

Dans un avenir proche, la diminution rapide du coût de séquençage et le développement d'appareils portables rendront routinière l'utilisation de la génomique en clinique, dans l'industrie et en ingénierie de l'environnement. Cela permettra le typage et la caractérisation rapide de bactéries résistantes aux antibiotiques, des avancés en bioremédiation ou encore la caractérisation et la classification rapide de tissus tumoraux ou de maladies génétiques. Pour exploiter ces nouvelles possibilités, il faudra être capable de gérer l'information qui sera produite de façon massive et délocalisée, de la mettre à jour et de l'intégrer. Alors qu'aujourd'hui les processus d'analyse automatique souffrent du manque d'harmonisation des banques de données et du manque d'outils suffisamment efficaces, il faudra définir des méthodologies d'analyse qui soient robustes et standardisées de façon à rendre les analyses moins dépendantes de l'expertise de l'utilisateur. Enfin, il faudra favoriser les interactions entre les bio-analystes, les utilisateurs de données (comme les cliniciens) et les laboratoires de recherche en bioinformatique. Le succès de ces démarches dépendra à la fois du développement de méthodes et d'infrastructures, mais aussi des efforts de recrutement, de communication et de formation. Pour que les avancées rapides de la génomique puissent être rapidement mises au service de la communauté, le lien entre les plate-formes de bioinformatique et la recherche méthodologique en biologie doit donc rester très étroit.

Les données de génomique ont vocation à être croisées avec des données fonctionnelles, d'épigénomique ou de criblage phénotypique pour augmenter la puissance des analyses et faciliter les études de biologie des systèmes. Il faudra donc mieux intégrer les milliers de banques de données en biologie contenant des informations à des niveaux très hétérogènes de détail, quantité et qualité. Techniquement, il faudra développer des outils permettant l'interopérabilité des différentes banques de données et assurer leur maintenance. Politiquement, il faudra valoriser le travail de recherche intégrative et la mise à disposition de données et d'outils.

II. Biologie structurale

La bioinformatique structurale est dédiée à l'étude et à la prédiction des structures des macromolécules biologiques comme les protéines, les acides nucléiques, les lipides ou les glycanes, de leurs interactions et de leurs complexes. Elle revêt deux aspects complémentaires. D'une part, des développements méthodologiques, dans lesquels les compétences interdisciplinaires en mathématiques, physique, chimie et informatique, nourries par les questions biologiques, sont essentielles et doivent être renforcées. D'autre part, un volet d'analyse, de structuration, de classification, d'exploitation et de valorisation de l'ensemble des données, souvent très hétérogènes, qui constituent le socle de connaissances sur lequel s'appuient les études fondamentales et prédictives, ainsi que la validation des nouvelles méthodologies. La France réalise dans ce domaine des développements novateurs et valorisés, notamment par la mise en place de serveurs souvent publiés dans des journaux spécialisés, la diffusion de logiciels et l'application à de très nombreuses thématiques biologiques.

Une des questions centrales de la bioinformatique structurale concerne l'étude des repliements des macromolécules, depuis la compréhension des mécanismes fondamentaux qui les gouvernent et la caractérisation de leurs briques fondamentales jusqu'à leur prédiction. Les études concernant les mécanismes de repliement ont très largement bénéficié ces dernières années de l'apport de techniques à très haute résolution temporelle. Par ailleurs, les développements méthodologiques se poursuivent dans les différentes voies de modélisation par homologie, par reconnaissance de repliement, ou *ab initio/de novo*. Ces approches facilitent la reconnaissance de parentés structurales entre protéines très éloignées sur le plan des séquences en acides aminés, ce qui est important pour poursuivre les efforts de modélisation à grande échelle et améliorer l'annotation fonctionnelle des génomes. Le couplage d'approches théoriques et expérimentales conduit également à des avancées significatives pour l'identification et la caractérisation des ARN non-codants et des structures désordonnées des protéines. La prédiction et la modélisation des interactions protéine-ligand et protéine-protéine est également un champ de recherche important, rejoignant pour partie le champ de la chémo-informatique.

La compréhension globale de la dynamique des complexes macromoléculaires nécessite le développement de méthodes d'intégration de données temporelles et spatiales multi-échelles, ainsi que de nouvelles méthodes de modélisation et de simulation. Toutefois, en parallèle, il convient de préserver, de développer et d'appliquer des expertises et des approches plus réductionnistes et de poursuivre des actions fondamentales et ciblées sur les acteurs même des complexes, en particulier pour la caractérisation et la modélisation des glucides, des lipides et des modifications post-traductionnelles, qui sont encore aujourd'hui insuffisamment explorées.

Grâce à l'augmentation des moyens de calcul et au développement de modèles physiques appropriés, la dynamique moléculaire connaît d'importantes évolutions, permettant de simuler des processus clés tels que le repliement, la liaison de ligands, ou les changements conformationnels de macromolécules. Ces avancées sont couplées au développement et à l'application de nouvelles méthodes d'exploration de l'espace énergétique, de calcul de l'énergie libre et de modèles simplifiés gros grain, permettant la mise en place d'approches multi-échelles. Mis en commun, ces développements autorisent un passage à l'échelle de la bioinformatique structurale, permettant la simulation et l'étude de très gros systèmes associant protéines, membranes, ADN et ARN, ou encore l'étude systématique par des méthodes de docking des interactions protéine-protéine. Ils laissent entrevoir à terme des interactions plus étroites avec le champ de la biologie des systèmes, même si plusieurs verrous, tant informatiques (exploration efficace sous contrainte de flexibilité) que biochimiques (calcul fiable de constantes d'association) restent encore à lever.

À l'interface avec la biologie structurale expérimentale, des méthodologies sont développées pour coupler les résultats obtenus à différentes résolutions par différentes techniques. Ces démarches seront d'autant plus essentielles que les informations structurales sont toujours plus nombreuses et variées, faisant écho au développement de nouvelles techniques comme les lasers à électrons libres. Les nombreux développements expérimentaux portant sur la cinétique suscitent également des développements spécifiques sur le plan de la bioinformatique. D'autres voient aussi le jour pour étudier la dynamique de complexes ou de structures particulières comme ceux des protéines intrinsèquement désordonnées ou des agrégats.

En conclusion, un effort conséquent de développement méthodologique à différents niveaux devrait permettre une intégration accrue de la bioinformatique structurale dans le contexte global et toujours évolutif du questionnement biologique, en lien étroit avec les expérimentateurs.

III. Biologie des systèmes

La biologie des systèmes est née au tournant des années 1990-2000 et a rapidement suscité un très fort intérêt dans la communauté biologique mais aussi aux interfaces de celle-ci avec la physique, l'informatique ou les mathématiques. En effet le programme scientifique de la biologie des systèmes répond à un besoin latent en biologie, à savoir disposer d'outils permettant de comprendre comment les systèmes vivants émergent comme des entités stables et structurées à partir de l'enchevêtrement complexe et dynamique des interactions moléculaires qui constituent leur substrat physico-chimique. L'apport de la biologie des systèmes, par rapport aux approches classiques de biologie moléculaire, de biologie cellulaire, de physiologie et de génétique, se situe dans la mobilisation systématique des outils de modélisation et de simulation dans l'analyse des systèmes biologiques, qui permettent de tester les hypothèses biologiques d'une manière formalisée, souvent quantitative, parfois même exhaustive. L'interdisciplinarité avec les mathématiques, l'informatique ou la physique en est donc une composante essentielle. En ce sens, on peut voir la biologie des systèmes comme une déclinaison biologique des « sciences de la complexité » à différentes échelles : cellule, tissu, organisme complet, voire éco-évolutive. De par sa nature intégrative et par la rationalisation, la quantification et l'identification de régulateurs clés des fonctions biologiques, la biologie des systèmes peut conduire à des applications importantes en recherche biomédicale. De même elle fournit un socle théorique essentiel à la biologie synthétique. En effet, produit du génie génétique et ayant comme objectif la réingénierie de systèmes biologiques, la biologie synthétique puise son inspiration dans des principes de design révélés par la biologie des systèmes

À partir de ce programme très général, la biologie des systèmes se décline en deux grands courants méthodologiques. D'une part les approches centrées sur les données, guidées par le développement rapide des techniques à haut débit et leur diversification à un très grand nombre de systèmes moléculaires et cellulaires en lien avec le développement du séquençage et des techniques d'imagerie cellulaire. Ces techniques permettent en effet d'acquérir directement un grand volume de données sur des sous-systèmes différents et, de plus en plus souvent, dans un contexte de cellule unique qui donne accès à la distribution des observables, voire même aux dynamiques stochastiques dans un contexte de molécule unique. Elles offrent donc à la biologie des systèmes une vue de plus en plus complète mais qui demande à être structurée par la modélisation, par exemple via l'apprentissage de modèles. D'autre part les approches centrées sur les systèmes qui ont pour objectif d'identifier les grands principes régissant la dynamique des systèmes biologiques en construisant des modèles computationnels de plus en plus précis de systèmes biologiques bien caractérisés dans la littérature. À de rares exceptions près, ces deux approches restent encore trop isolées l'une de l'autre alors que l'objectif premier de la biologie des systèmes devrait être de les réconcilier. La faute en est probablement au morcellement disciplinaire (l'approche données-centrée venant souvent des laboratoires de biologie tandis que l'approche système-centrée est souvent issue des modélisateurs) mais aussi à l'insuffisance de formations transversales. La biologie des systèmes demande en effet que les chercheurs soient à la fois rompus aux techniques expérimentales mais aussi aux méthodes et outils d'analyse de données et à la modélisation mathématique. Un bagage que peu de formations dispensent à l'heure actuelle.

Malgré son très jeune âge – ou peut-être en raison de celui-ci – la biologie des systèmes est en évolution permanente et rapide. On est ainsi passé en quelques années des approches cartographiques en « grands réseaux », où les systèmes biologiques étaient décrits par les interconnexions statiques entre leurs composants moléculaires, à la prise en compte des dynamiques temporelles puis spatio-temporelles de ces mêmes réseaux au moyen de for-

malismes très divers, allant des systèmes dynamiques aux méthodes informatiques formelles en passant par la simulation individu-centrée ou la fouille de graphes statiques ou dynamiques. Les recherches théoriques parallèles qui concernent l'acquisition et le traitement de données à haut débit et d'imagerie cellulaire et tissulaire pourront se révéler cruciales pour le développement de ce domaine.

La France a des atouts en biologie des systèmes, avec une forte tradition en modélisation, en théorie du contrôle et en méthodes formelles appliquées à la biologie. Cependant, force est de constater que l'investissement individuel et collectif des chercheurs en biologie des systèmes n'a pas été soutenu à la hauteur de ce qu'il a été aux États-Unis, au Japon, en Allemagne et en Grande-Bretagne. En conséquence, la France manque de structures d'envergure dans ce domaine et joue un rôle moins important qu'elle ne le devrait. Il est donc stratégique de développer et de soutenir les initiatives interdisciplinaires qui intègrent des approches expérimentales et des approches de modélisation et d'analyse de données en veillant à ce qu'elles se développent au sein d'une même équipe ou en collaboration très étroite entre équipes complémentaires.

IV. Neurobiologie, cognition

Dans les dix dernières années, les neurosciences se sont beaucoup développées notamment dans leur côté le plus fondamental représenté par l'étude cellulaire des neurones et les neurosciences computationnelles mais aussi dans les domaines plus intégrés qui constituent le point de départ d'une interface avec la médecine. Dans ce contexte, l'interface s'établit d'abord au niveau de l'analyse de données biologiques qui s'adresse aussi bien aux images qu'aux enregistrements électrophysiologiques puis au niveau de la modélisation qui fait appel plus ou moins directement aux statistiques et à l'informatique. Les neurosciences sont, en France, réparties dans de nombreux instituts et unités. Les chercheurs dont le domaine principal d'investigation est la modélisation et les neurosciences computationnelles sont rarement dans les mêmes équipes que les expérimentateurs. Cependant plusieurs unités développent depuis un certain temps et avec beaucoup de succès des interactions interdisciplinaires proches au quotidien.

La dernière décennie a vu une explosion des techniques d'imagerie au sens large, impliquant la collecte de grandes masses de données. Les défis posés sont l'amélioration de la qualité et de la pertinence des données recueillies ainsi que leur analyse quand il s'agit de grandes masses de données complexes. L'imagerie cellulaire des neurones à proprement parler et en particulier l'imagerie calcique sont des méthodes d'études qui sont utilisées en neurosciences fondamentales et qui contribuent à l'exploration du système nerveux. Ces techniques et conjointement les méthodes d'analyse qui leur correspondent sont en plein essor et devraient rapidement permettre une compréhension du système nerveux *in situ*. Parallèlement, les neurosciences des invertébrés, en particulier des insectes, connaissent une forte progression. Leur grande disponibilité, la mise à disposition de banques de mutants et la possibilité de réaliser des enregistrements électrophysiologiques *in vivo* sur plusieurs jours, font de l'insecte un remarquable modèle d'investigation *in vivo* où l'enregistrement de l'activité neuronale peut être facilement couplé à une tâche olfactive par exemple. Ainsi l'obtention de grandes quantités de données sous forme de trains de potentiels d'action (encore plus remarquable en utilisant des électrodes multisites) requiert de nouvelles méthodes pour les trier et classifier. En conséquence, le développement de méthodes statistiques dédiées à la modélisation et à l'analyse des données spécifiques aux neurosciences doit être poursuivi et accéléré afin d'améliorer la pertinence des modèles proposés. Il est à noter que l'enregistrement de potentiels d'action seuls ne permettra pas d'approcher le phénomène d'inhibition, clé de la compréhension des patrons de décharge obtenus *in vivo*. Seule l'utilisation de sondes

fluorescentes sensibles au potentiel de membrane devrait être capable de fournir de telles informations à grande échelle et dans un système intégré. Les sondes aujourd'hui disponibles ne proposent pas une sensibilité suffisante pour analyser correctement ces phénomènes mais nous pouvons d'ores et déjà prévoir que ce sera chose faite dans les années à venir.

Les structures de connectivité au sein d'un réseau de neurones ou à plus grande échelle entre des aires cérébrales différentes, ainsi que leur dynamique, apparaissent comme des éléments essentiels pour la compréhension du fonctionnement du cerveau. La notion de connectivité s'est considérablement affinée ces dernières années. L'inférence des structures de connectivité est en plein essor. À l'échelle des réseaux de neurones, des modèles de processus statistiques utilisés depuis longtemps comme les processus de Hawkes sont réapparus permettant de mettre au point des méthodes puissantes d'analyse de connectivité. À une échelle plus grande, des techniques statistiques utilisant des modélisations linéaires ou utilisant la théorie de l'information permettent d'inférer la circulation de l'information entre des aires cérébrales. Les structures inférées sont en général analysées à l'aune de la théorie des graphes, et particulièrement de la théorie des graphes de terrain (complex networks). Des informations importantes sur le cerveau, par exemple sur la résilience des systèmes neuronaux, peuvent être tirées de la topologie des graphes de connectivité inférés.

L'éthologie est la science qui étudie le comportement des animaux ainsi que ses déterminants physiologiques et environnementaux. L'éthologie est une science transversale qui chevauche des disciplines variées comme la physiologie, l'écologie, la sociologie, la psychologie sociale, les neurosciences... L'éthologie tente aujourd'hui d'expliciter le comportement résultant des interactions entre les constituants du vivant à différentes échelles. Les être vivants sont des systèmes complexes intégrés dans un environnement au moins aussi complexe et dont les comportements sont impossibles à prévoir à partir de la seule connaissance des propriétés de leurs constituants. Le principal enjeu est de comprendre comment les propriétés observées à l'échelle d'un système biologique (exemple : une société animale) émergent d'un ensemble complexe d'interactions entre ses différents éléments (individus). À chaque niveau d'organisation, un très grand nombre de constituants interagissent de manière non-linéaire et permettent au système de s'auto-organiser spontanément. Pour comprendre ces phénomènes, l'éthologie doit adopter une démarche itérative et intégrative en combinant des approches expérimentales et théoriques dans lesquelles la modélisation mathématique et la simulation jouent un rôle central. Ces modèles sont construits à partir des lois établies à l'échelle des constituants et permettent ensuite une analyse des propriétés résultant de leurs interactions. Les simulations numériques de ces modèles permettent en particulier de déterminer les effets qualitatifs et quantitatifs de chaque paramètre du comportement des constituants sur la dynamique et les caractéristiques spatiales et/ou temporelles des phénomènes produits à l'échelle supérieure. En associant étroitement expérience et modélisation, il est possible de comprendre un grand nombre de phénomènes comportementaux à différents niveaux d'organisation (mouvement de foule, morphogénèse de nid d'insecte, réseaux sociaux, etc.). Ainsi, le champ des observations en éthologie est passé de l'étude d'objets biologiques simples à des systèmes complexes d'entités en interaction. Face à cette situation, l'éthologie et ses champs d'application doivent continuer à s'ouvrir à des disciplines éloignées telles que la chimie et la physique pour le développement et la conception de nouveaux instruments d'observation ; les statistiques pour l'intégration et la représentation de données hétérogènes et de grande dimension ; l'informatique pour la gestion de bases de données ; les mathématiques pour modéliser des phénomènes complexes et construire des outils prédictifs. De grands centres de recherche nationaux se sont développés pour répondre à ces nouveaux défis.

V. Écologie, évolution

Les domaines de l'écologie et de la biologie évolutive ont subi une transformation radicale dans les deux dernières décennies. Premièrement, l'acquisition croissante des données de génomique, métagénomique, protéomique et métabolomique a transformé notre perception de la biodiversité et notre capacité à comprendre les mécanismes évolutifs et écologiques. Deuxièmement, l'observation des (et l'expérimentation sur les) systèmes écologiques a atteint des échelles de temps, d'espace et de complexité sans précédent grâce à des avancées technologiques majeures. Troisièmement, les enjeux scientifiques et sociétaux posés par les changements globaux posent des défis cruciaux et urgents en termes de compréhension et de prédiction des phénomènes d'adaptation et de fonctionnement des écosystèmes. Ces transformations radicales appellent de plus en plus une vision pluridisciplinaire et les approches de modélisation et d'analyse de données issues des mathématiques, de l'informatique et de la physique sont devenues cruciales. Allant du gène à l'écosystème, plusieurs avancées majeures en écologie et évolution se profilent à l'interface avec les autres disciplines.

Renforcer le dialogue entre la génétique et l'écologie serait l'occasion de proposer des approches quantitatives permettant de passer de l'échelle des populations à celle des communautés. Dans ce contexte, il faudra développer le transfert vers l'écologie des méthodes mathématiques et statistiques utilisées en génétique. Réciproquement, l'écologie pourra aider à théoriser et interpréter les données génomiques et métagénomiques. Les dernières années ont vu le développement d'approches basées sur l'écologie en santé humaine, comme l'utilisation de virus pour contrôler les bactéries résistantes aux antibiotiques ou l'utilisation de commensaux pour empêcher la colonisation par des pathogènes. Les approches écologiques trouvent aussi des applications importantes dans l'agro-alimentaire.

Les études d'évolution expérimentale couplées avec des analyses génomiques connaissent actuellement de beaux succès et ouvrent la possibilité de mieux comprendre les relations génotype-phénotype ainsi que leur évolution. Elles ont permis de tester de nombreux modèles de génétique des populations et de mieux comprendre leurs limites. De plus, l'accroissement de la complexité de ces expériences permet de tester des modèles écologiques *in vitro* sur des communautés de microorganismes, leurs parasites et leurs prédateurs. Ces approches ouvrent encore plus le domaine de l'écologie évolutive à l'expérimentation dans un cadre contrôlé permettant la paramétrisation des modèles et la manipulation expérimentale des interactions. De plus, une liaison forte est en train de s'établir entre l'évolution expérimentale et l'évolution *in silico* du type vie artificielle, permettant la modélisation de phénomènes complexes pour lesquels les outils de modélisation classiques sont peu adaptés.

Un autre domaine de l'écologie faisant appel à des compétences interdisciplinaires concerne le rapprochement entre la modélisation en écologie et la phylogénie. La possibilité de reconstruire des phylogénies à grande échelle est une opportunité pour mieux comprendre comment s'est façonnée la biodiversité au cours du temps. La phylogénie peut fournir des informations importantes sur les processus écologiques passés, la co-évolution et les phénomènes de spéciation. Pour développer l'interaction entre la phylogénie et l'écologie il sera important d'intéresser la très riche communauté française développant des approches mathématiques et informatiques en reconstruction phylogénétique au développement de méthodes phylogénétiques comparatives et leurs applications en écologie et évolution. Le développement de ces approches permettrait d'améliorer notre compréhension des conséquences à long terme de l'adaptation sur la diversité phénotypique et spécifique et de mieux comprendre les processus à l'origine de l'assemblage des communautés.

Parallèlement à l'utilisation des modèles mathématiques, il y a un réel besoin de modèles

couplant phénomènes physiques et biologiques pour comprendre et prédire l'impact potentiel des changements climatiques sur la biosphère. La compréhension des raisons de la crise actuelle de la biodiversité et du réchauffement climatique et la définition des mesures à prendre pour minimiser leur impact sont des problèmes dont la résolution sollicitera plusieurs disciplines. De manière plus générale, avec la complexification des processus évolutifs et écologiques à modéliser, les approches mathématiques doivent être complétées par des approches relevant de la physique ou des sciences computationnelles.

VI. Interface physique

Les approches de modélisation à l'interface de la biologie avec la physique permettent de développer des concepts structurants et d'obtenir des descriptions du fonctionnement des systèmes biologiques à différentes échelles d'organisation : depuis le niveau moléculaire, jusqu'aux machines cellulaires complexes, tissus, organes, organismes et écosystèmes. La physique biologique, la biologie structurale et la biologie des systèmes ont comme objectif commun l'étude des mécanismes du vivant par des modèles. On peut naturellement s'attendre à une convergence des méthodes, questions et problématiques de ces domaines scientifiques.

La physique statistique a produit des contributions importantes en biophysique, biochimie et plus récemment en relation avec l'analyse des données génomiques. Des modèles de la physique statistique, tels que les modèles d'Ising, Potts, gaz sur réseau, marches aléatoires et modèles de polymères ont permis d'analyser le fonctionnement des macromolécules à l'origine de couplages bio-mécano-chimiques dans des processus tels que l'adhésion, la motilité, la division cellulaire, la différentiation cellulaire ou l'expression des gènes. Ces études ont conduit à de nouveaux paradigmes tels que la relation entre contraintes mécaniques et expression génique. Des modèles mécaniques d'organes et de tissus cellulaires trouvent des applications en médecine, notamment dans l'étude des cancers où ils soulignent l'importance de la mécanique et de la physico-chimie dans les phénomènes d'invasion et de prolifération tumorale. Un effort particulier est nécessaire pour réaliser le passage à l'échelle des modèles moléculaires et cellulaires aux modèles tissulaires et d'organes. L'enjeu est de construire des modèles réalistes, physiquement cohérents et qui contiennent suffisamment de détails pour intégrer des données génomiques et biophysiques. Les modèles de la physique non-linéaire, tels que les solitons, ont permis d'aborder des problèmes biologiques divers allant de la dénaturation de l'ADN à la formation de motifs en morphogènese et la dynamique spatiale des populations en écologie. À travers le concept général de criticalité auto-organisée, la physique non-linéaire a inspiré des recherches en neurosciences, en conduisant à des interprétations nouvelles des signatures caractéristiques de l'activité des neurones avec des possibles applications en diagnostic. Finalement, et ce qui n'est pas le moins important, la modélisation biophysique ne peut pas être dissociée de l'expérimentation et de l'analyse de données. Les programmes de recherche à la frontière entre la physique et la biologie profitent ainsi des techniques de pointe en biophysique de la molécule unique, mesures de force mécanique, imagerie multimodale.

En France, l'interface entre la physique et la biologie est bien développée et structurée grâce à une forte tradition en physico-chimie et à une nouvelle génération de physiciens attirés par la complexité des systèmes biologiques. Le lien entre la modélisation et l'expérimentation biophysique est facilité par l'existence de nombreuses équipes reconnues au plus haut niveau international et de bonnes pratiques de collaboration au niveau de chercheurs sur le terrain. Cependant, la France est en retard dans l'élaboration et l'enseignement des nouveaux concepts physiques issus des avancées expérimentales en génomique et en biophysique. La compétition internationale dans ce domaine demande aux chercheurs

non seulement une solide formation en physique théorique mais aussi des connaissances étendues en biologie. En outre, des synergies sont nécessaires entre les physiciens, les bioinformaticiens et les modélisateurs impliqués en biologie des systèmes et en biologie structurale pour relever les défis de la modélisation à grande échelle.

VII. Interface mathématiques

La modélisation mathématique en sciences du vivant doit répondre à des questions spécifiques pour rendre possible une compréhension quantitative des phénomènes mais aussi permettre l'émergence de concepts plus généraux. Les études mathématiques sont souvent effectuées sur des systèmes simplifiés qui font ressortir les concepts fondamentaux tandis que les approches par simulation peuvent inclure des aspects plus détaillés et spécifiques de chaque système biologique. Le très bon niveau de l'école mathématique française a permis l'émergence d'un grand nombre de petites équipes de haut niveau internationalement reconnues à l'interface entre mathématiques et sciences du vivant. Néanmoins, ce succès repose trop souvent sur des initiatives individuelles de chercheurs n'ayant pas à l'origine été formés pour cela. La création d'un centre de recherche ou de rencontre dédié serait extrêmement utile pour la formation avancée et la recherche.

Traditionnellement la biologie quantitative décrit surtout les états stationnaires – ceux-ci étant les plus facilement observables et quantifiables expérimentalement. Grâce aux évolutions récentes des techniques d'acquisition et de traitement de données, il devient possible d'observer en détail la dynamique de nombreux systèmes biologiques. Par exemple, les nouvelles techniques d'acquisition d'images en microscopie et de marquage de protéines permettent de faire des films avec une grande résolution spatiale et temporelle au niveau d'une cellule, d'un tissu/organe ou même d'une région du corps. Ceci ouvre la possibilité d'élaborer et paramétrer des modèles dynamiques réalistes, qui sont beaucoup plus riches du point de vue mathématique.

La construction de modèles mathématiques *in silico* plus réalistes prenant en compte des réseaux de régulation et la physique (en particulier la mécanique) des systèmes biologiques répond à une demande forte de la part des sciences du vivant. De nombreuses équipes mathématiques y travaillent actuellement en France (en étroite collaboration avec des biophysiciens) sur des thématiques très diverses parmi lesquelles l'hémodynamique, la croissance de tissus et de tumeurs, la dynamique de populations et les neurosciences.

La mise en place de modèles hybrides permettant de marier des descriptions discrètes au niveau des cellules ou des individus avec des modèles continus au niveau de la population ou du milieu environnant suscite un effort important de la communauté. Ce type d'approche permet de profiter simultanément des avantages des modèles discrets ou à base d'agents pour une description détaillée d'une petite région ou groupe de cellules, avec le moindre coût computationnel des modèles continus qui permettent d'avoir une description réaliste sur des échelles plus grandes. Les mathématiciens utilisent déjà un grand nombre d'outils de changement d'échelle qui permettent de faire le lien entre des comportements microscopiques stochastiques ou déterministes et des modèles continus au niveau macroscopique comme les méthodes d'homogénéisation, les théorèmes limites des processus stochastiques, les modèles de champ moyen et cinétiques pour des systèmes de particules en interaction. L'étude des systèmes biologiques incitera leur développement et l'émergence de nouvelles approches plus adaptées à la problématique du vivant où les systèmes ont souvent une réponse complexe.

Un autre domaine en grand essor en ce moment et qui devra s'intensifier dans le futur, concerne les applications de la théorie du contrôle dans le contexte du vivant. Elles ont un intérêt large, mais sont particulièrement

pertinentes dans le contexte médical où il est question d'amener le patient vers un état souhaité – la guérison – de façon optimale. On cherchera par exemple à minimiser la durée ou les effets secondaires des traitements qui sont mesurés par la variable de contrôle. Des travaux récents sur l'optimisation de doses de médicaments ou la combinaison de différents traitements pour le cancer se basent sur des modèles très simplifiés, mais sont prometteurs et devront pouvoir devenir plus réalistes dans un futur proche.

Une fois les modèles validés, leur étude et la simulation *in silico* permettent de guider des choix biologiques et de réduire le nombre d'expériences à réaliser. Dans l'ère des grands jeux de données, en parallèle avec des avancées informatiques sur le stockage de grandes masses de données, la modélisation mathématique aura un rôle essentiel dans la réduction de la quantité d'information à stocker en la limitant à un petit nombre de paramètres significatifs qui peuvent êtres estimés par différentes approches.

Les nouvelles capacités de recueil et de stockage des données provoque un changement de paradigme en nécessitant de nouvelles compétences pour les statisticiens comme l'analyse numérique, la gestion informatique de grandes bases de données ou l'utilisation de méthodes séquentielles. Il s'agit par exemple de développer des méthodes permettant d'analyser des données structurées en réseaux, de classer des données en grande dimension ou de traiter des données hétérogènes. Nombre de données biologiques contiennent plus de variables que d'individus et les méthodes d'estimation ont été adaptées en utilisant des techniques dites de régularisation qui permettent de pénaliser la complexité des modèles. Le transfert de ces techniques issues du machine learning à la biologie fait l'objet d'un effort de recherche important notamment en bioinformatique. Les approches bayésiennes ont elles aussi connu d'importants développements, basées en partie sur des simulations stochastiques de type Monte Carlo (algorithmes MCMC, ABC).

Ce qui est remarquable c'est que non seulement la modélisation et l'analyse de données biologiques bénéficient des avancées méthodologiques apportées par les mathématiques mais la biologie a aussi ouvert de nouvelles perspectives de recherche dans un grand nombre de domaines des mathématiques.

VIII. Interface sciences de l'information

L'interface entre sciences de l'information et sciences du vivant s'est révélée, depuis déjà quelques décennies, un exemple particulièrement marquant de fertilisation croisée réussie. D'une part, des algorithmes efficaces ont accompagné la biologie dans l'ère du haut débit. D'autre part, les organismes biologiques évoluent par sélection naturelle, un processus d'optimisation, dont la mémoire est en partie gravée dans les génomes, eux-mêmes un exemple d'information digitale. En conséquence, l'interface biologie/sciences de l'information a fait émerger de nouveaux modèles de calculs bio-inspirés. Il en résulte un dialogue souvent fructueux, dans lequel le rôle des bioinformaticiens ne se limite pas à la proposition d'une solution technique efficace, mais où ils participent à la conception d'expériences de validation, voire au choix des questions biologiques.

Les contributions de l'informatique au niveau de l'algorithmique sont bien représentées au niveau national en optimisation combinatoire, par exemple en phylogénie, en génomique et en biologie structurale. Dans ces domaines, le traitement de données volumineuses et de qualité hétérogène nécessite des structures de données compressées et des algorithmes efficaces et résilients aux erreurs. Les problématiques biologiques nécessitent souvent des approches hybrides, intégrant simultanément plusieurs types de données dans un schéma algorithmique joint, pouvant

mêler inférence de modèle et optimisation. L'approche réductionniste traditionnellement utilisée pour la conception des algorithmes est alors mise à mal. Les méthodes issues de la recherche opérationnelle pourraient apporter des solutions à ces problèmes.

Les développements récents en biologie confirment la nécessité de développer de nouvelles méthodes et outils en sciences de l'information, afin de faire face à l'explosion des volumes de données issues de la pratique quotidienne des sciences du vivant. Cela concerne l'organisation des données sur des ontologies spécifiques, ainsi que des workflows permettant une rationalisation et une modularisation des méthodologies d'analyse. Des contributions méthodologiques importantes sont aussi attendues dans le domaine de l'apprentissage et de l'extraction de connaissances. Elles concernent entre autre la classification automatique, domaine où les développements informatiques et bioinformatiques se nourrissent mutuellement car les études de cas biologiques sont désormais intégrées aux procédures d'évaluations de nouvelles contributions disciplinaires. Outre les problèmes classiques, mais toujours d'actualité, d'apprentissage et inférence dans des espaces de très grande dimension, la confrontation à des volumes de données distribuées, et produits en continu, constitue un défi majeur. Ces problématiques sont actuellement au cœur de nombreux travaux et concepts dans les communautés de l'automatique, du traitement du signal et des images, de l'apprentissage artificiel, et devraient bientôt alimenter la communauté bioinformatique.

Enfin, parallèlement à ces aspects liés à l'analyse de données haut-débit et permettant de développer des modèles à partir des données, l'interface entre informatique et sciences du vivant présente une facette de modélisation « a priori » dans laquelle ce sont les processus qui sont modélisés informatiquement au moyen de différentes approches telles que les automates cellulaires, les réseaux de Petri, les modèles individu-centrés, ou les langages formels. L'objectif de la modélisation n'est plus ici d'extraire du sens à partir d'un ensemble de données mais d'inférer le régime de fonctionnement le plus probable d'un système biologique ou de révéler des liens de causalité qu'il serait impossible d'identifier spontanément du fait de la complexité du système ou de ses non-linéarités. Ces approches se développent très fortement dans le courant des sciences computationnelles en lien avec le développement de la biologie des systèmes (voir section III). Elles permettent de développer un cercle vertueux entre modélisation et expérimentation en proposant, par la modélisation, des hypothèses qui peuvent ensuite être mises à l'épreuve expérimentalement. Cependant, pour rester vertueux, ce cercle doit impérativement intégrer le plus étroitement possible les deux approches ce qui ne peut se faire que dans des groupes fortement interdisciplinaires. Cette nécessité contraste fortement avec la volonté de créer de grosses entités de recherche regroupant tous les chercheurs d'une même discipline dans une logique de site (voir section XI).

IX. Logiciels et calcul

La production logicielle directement issue de la communauté scientifique joue un rôle de tout premier plan dans la modélisation et l'analyse des systèmes biologiques. Elle assure une partie de la continuité entre les deux versants que sont la recherche fondamentale en biologie et les développements méthodologiques et numériques. Le succès d'un logiciel est fonction de l'avancée méthodologique qu'il porte, mais aussi de son ergonomie, de sa disponibilité et de sa capacité à suivre l'avancée de son domaine d'application. Pour les développeurs de méthodes, le fait que leurs approches puissent être utilisées à grande échelle constitue un test grandeur nature pouvant mettre à jour forces et faiblesses. Ainsi, la recherche en biologie et les développements méthodologiques s'enrichissent et se fertilisent mutuellement via les logiciels qui agissent comme des vecteurs de communication.

Au sein des unités CNRS, le développement de logiciels pour analyser et simuler des données biologiques compte des succès comme ClustalX, GenePop ou PhyML qui ont atteint la dizaine de milliers de citations et d'autres comme SeaView/Phylo_win, T-Coffee, EEGLAB ou APE qui sont des références dans leurs domaines. De façon intéressante, la plupart des ces logiciels concernent les domaines de l'analyse de séquences et de la biologie évolutive. La communauté française dans d'autres domaines de la bioinformatique a produit des avancées méthodologiques importantes mais qui n'ont peut-être pas été suffisamment exploitées en partie en raison des difficultés rencontrées par les chercheurs impliqués dans un développement logiciel. Nous proposons ici des pistes pour développer et rationaliser le développement de logiciels scientifiques et la capacité de calcul.

1. Développer des logiciels libres. Le logiciel libre constitue une forme fructueuse de travail collaboratif qui favorise la reproductibilité des recherches et permet d'exposer les erreurs potentielles.

2. Promouvoir l'accès de la biologie aux infrastructures informatiques. Vu les volumes de données générés en biologie ainsi que le besoin croissant en puissance de calcul, il est de moins en moins raisonnable de rester sur un modèle de stockage et de calcul exclusivement local.

3. Développer et maintenir les bases de données. En effet, elles conditionnent la qualité des analyses bioinformatiques.

4. Évaluer plus favorablement les activités liées au logiciel. Si le développement logiciel ne relève pas de la science à proprement parler, il est au cœur de l'interdisciplinarité car il joue un rôle clé dans la diffusion des méthodes parmi les biologistes.

5. Fournir des moyens techniques mais aussi humains pour maintenir les logiciels.

X. La communauté

Les domaines interdisciplinaires sont flous et mouvants avant d'éventuellement devenir des disciplines à part entière. Cela oblige à un effort constant d'animation et de structuration d'une communauté scientifique dont la définition reste inévitablement imprécise. Ainsi, alors que des chercheurs restent attachés à leurs communautés d'origine et utilisent la biologie comme source d'inspiration pour des questions disciplinaires, d'autres prennent le risque de les abandonner. Les chercheurs qui restent au carrefour de différentes communautés peuvent être confrontés à des problèmes de communication, de reconnaissance et d'évaluation.

Plusieurs sociétés savantes aident à structurer la communauté de la CID51. La Société Française de Bioinformatique (SFBI) organise un congrès qui réunit entre 350 et 500 chercheurs tous les ans (JOBIM) et gère une liste de diffusion qui atteint plus de 5000 personnes en France et à l'étranger. La SFBI regroupe surtout des chercheurs autour de l'interaction entre informatique et biologie, en incluant la génomique, l'évolution, la biologie structurale et la biologie des systèmes. Les structuralistes se retrouvent aussi au sein du Groupe Graphisme et Modélisation Moléculaire. La Société Francophone de Biologie Théorique organise des séminaires annuels et des congrès internationaux francophones autour des relations entre les mathématiques et la biologie. La Société Française de Biophysique est très liée aux recherches sur les structures des molécules et la biochimie théorique et organise tous les deux ans un congrès national. Il existe aussi une communauté riche de modélisateurs qui se retrouvent dans des colloques et des écoles telles que l'école thématique interdisciplinaire du CNRS Berder-Porquerolles, ainsi qu'une communauté centrée sur la biologie des systèmes et la biologie de synthèse qui se réunit autour d'une école de chercheurs depuis plus de dix ans.

Les trois principaux outils du CNRS pour structurer l'interdisciplinarité sont les CID, la Mission pour l'Interdisciplinarité et les GdR. La deuxième finance de courts projets interdisciplinaires. Cela permet à de jeunes chercheurs de coordonner leur premier projet de recherche et facilite la mise en place de nouvelles collaborations, même si les montants accordés sont souvent trop faibles pour avoir un impact important en termes de recherche. Les GdR facilitent la discussion entre chercheurs de différentes disciplines, la création d'écoles thématiques et de réseaux de collaboration. Il y a des nombreux GdRs dans le champ thématique de la CID51 sur des domaines comme l'imagerie et le traitement du signal, la bioinformatique moléculaire, l'écologie statistique, la biophysique, et très récemment la biologie des systèmes.

XI. Verrous institutionnels

Même si l'interdisciplinarité en biologie est souvent présentée comme prioritaire dans les discours de politique scientifique, beaucoup de verrous institutionnels freinent son développement. Ces verrous sont le résultat de contraintes diverses. Premièrement, la structuration de l'activité de recherche en universités, instituts ou UFRs disciplinaires, complique l'émergence de domaines interdisciplinaires. Cette structuration impose aussi des contraintes de recrutement, d'affectation ou d'évaluation qui défavorisent les chercheurs interdisciplinaires car ils se trouvent en général face à des arbitrages et des jurys aux compétences centrées sur une seule discipline. Le CNRS a l'originalité d'avoir développé les CIDs pour le recrutement et la co-évaluation des chercheurs interdisciplinaires. Cela a permis une intégration forte de compétences autour de la modélisation et l'analyse de données en biologie, sans pour autant résoudre entièrement les problèmes de recrutement. En effet, le nombre de candidats dans la CID51 est de l'ordre de 60 par poste non colorié (ouvert à plusieurs instituts disciplinaires), plus que la moyenne au CNRS. À l'université, la situation est encore plus délicate pour les enseignants-chercheurs dans des domaines interdisciplinaires car les postes dépendent des besoins en enseignement des UFRs et ont donc tendance à favoriser l'aspect disciplinaire des profils. La division du CNU et des UFRs en domaines strictement disciplinaires rendent aussi difficile la promotion des enseignants-chercheurs et le développement de laboratoires interdisciplinaires. La création des très grandes unités CNRS centrées autour d'une discipline pourra rendre la situation encore plus délicate s'il n'y a pas un effort d'évaluation spécifique des équipes interdisciplinaires. Bien évidemment, la situation n'est guère plus favorable dans les nombreux établissements français de recherche mono-disciplinaires.

Les financements sur projet ainsi que le recrutement de doctorants peuvent aussi être plus complexes pour les chercheurs à l'interface par manque de soutien des commissions d'évaluation, typiquement disciplinaires. Dans l'environnement actuel, avec des taux très faibles d'acceptation de dossiers, les commissions d'évaluation, souvent peu à même de s'assurer de la qualité des projets interdisciplinaires, ne prennent pas le risque de les soutenir. Les différents modes de fonctionnement de chaque communauté compliquent aussi la structuration de la recherche et la publication des résultats. L'influence grandissante des financements par projet et l'importance du prestige de la liste de publications pour l'obtention de ces financements risquent donc de freiner considérablement le développement de l'interface entre la biologie et d'autres domaines de recherche.

XII. Constats et recommandations

La France a des atouts importants dans les domaines de recherche associés à la bioinformatique et à la modélisation en biologie. La

formation des étudiants aux mathématiques y est plus poussée qu'ailleurs et la robustesse du système de recherche français aux effets de mode devraient faciliter l'établissement de projets interdisciplinaires ambitieux. Pour favoriser ces développements, nous proposons une série de mesures destinées à contrer les verrous mentionnés ci-dessus :

Introduction à l'interdisciplinarité. Les thématiques interdisciplinaires devraient être systématiquement introduites dès le niveau licence en mathématiques, informatique et physique pour susciter des vocations et permettre d'acquérir une double culture très tôt. Il faudra aussi proposer aux biologistes une offre plus complète de formations interdisciplinaires pour approfondir les connaissances biologiques tout en assurant une bonne formation aux autres disciplines et aux techniques de modélisation des systèmes vivants.

Formation doctorale. Augmenter le nombre de thèses en co-encadrement et de thèses interdisciplinaires permettrait aux étudiants d'être confrontés à différentes pratiques de la science dès le doctorat. En outre, le regroupement des équipes d'accueil en écoles doctorales disciplinaires rend difficile la co-existence d'étudiants et encadrants de différentes disciplines dans les mêmes équipes. La création d'écoles doctorales (ou de programmes doctoraux) permettant l'encadrement de thèses interdisciplinaires sans que les encadrants sortent de leur école doctorale disciplinaire (à l'image de l'école doctorale Frontières du Vivant) faciliterait la formation d'étudiants à l'interface des disciplines.

Recrutement. Le recrutement est un élément central pour le développement des nouveaux domaines à l'interface et la pénurie actuelle de postes risque de freiner sérieusement leur essor. Les CIDs permettent au CNRS de réaliser des recrutements sur la base de jurys interdisciplinaires et certaines universités créent aussi des comités de sélection appropriés. Il est important que la pénurie de postes ne se traduise pas par un recentrage des unités de recherche et des UFRs uniquement sur leur cœur de compétence. Cela serait un mauvais signal pour les développements interdisciplinaires qui sont pourtant à l'origine de beaucoup d'avancées récentes.

Affiliations. Étant le seul établissement national réunissant toutes les disciplines scientifiques, le CNRS se doit de jouer un rôle majeur dans le développement de la modélisation et l'analyse de données en biologie. Néanmoins, même au sein du CNRS plusieurs obstacles compliquent le développement de projets interdisciplinaires. Le rattachement des laboratoires aux instituts complique la mobilité thématique des chercheurs qui se retrouvent dans une unité dont les ressources et les ITA sont associés à l'institut d'affiliation principale et dont les sections d'évaluation ne coïncident pas avec les leurs. Cela pourrait se résoudre en développant la double affiliation des équipes et des chercheurs au sein des UMR. De tels systèmes ont permis aux États-unis et au Royaume Uni de développer beaucoup plus vite les domaines interdisciplinaires en biologie, par exemple en biologie des systèmes. De même, un enseignant-chercheur devrait pouvoir, s'il le souhaite, appartenir à plusieurs UFR et être évalué par plusieurs sections du CNU. Des sections interdisciplinaires dans le CNU, à l'image des CID, favoriseraient aussi une évaluation correcte des enseignants-chercheurs à l'interface. À l'image des équipes-projets INRIA hébergées dans des unités CNRS, la création d'équipes préférentiellement rattachées à un institut au sein d'unités rattachées à un autre institut rendrait les disciplines plus perméables entre elles.

Évaluation. Les équipes interdisciplinaires sont rarement évaluées par des jurys couvrant l'ensemble de leurs domaines de compétence et cela nuit à la qualité de l'évaluation. Il faudra que les décisions de composition de ces jurys prennent en compte l'existence de ces équipes au sein de structures plus vastes. L'HCERES devrait favoriser spécifiquement dans ses évaluations les efforts d'interdisciplinarité.

Financement. Il y a actuellement peu de sources de financement favorisant spécifiquement l'interdisciplinarité. Des appels à projets pour des montants plus importants que les PEPS permettraient aux réseaux interdisciplinaires de se constituer avant d'essayer des appels d'offre ANR ou autres. La création de

jurys interdisciplinaires à l'ANR favoriserait le développement de projets de qualité à l'interface. Des systèmes de financement du genre ATIPE ou ANR Jeunes Chercheurs pour des projets fortement interdisciplinaires avec des jurys adaptés constitueraient aussi une incitation importante.

Structuration de la communauté. Beaucoup d'équipes interdisciplinaires sont de petite taille et se retrouvent isolées au sein d'unités fortement disciplinaires. La création de structures d'échanges entre disciplines pourrait passer par la création de lieux de rencontre institutionnels et des possibilités d'immersion dans des laboratoires d'autres disciplines sur des périodes longues. Dans ce contexte, la création de groupements de recherche spécifiquement interdisciplinaires, donc affiliés à plusieurs instituts ou à la direction du CNRS, favoriserait la structuration de la communauté. La création d'un répertoire national accessible avec l'information et les contacts des GdR existants augmenterait aussi leur capacité à regrouper la communauté. De même les possibilités de délégation au CNRS pour les enseignants-chercheurs d'autres domaines scientifiques devraient être développées (et évaluées par des comités interdisciplinaires).

Relations entre plate-formes, biologistes et chercheurs en modélisation. La dernière décennie a vu la création de très nombreuses plate-formes de bioinformatique en France. Cela n'a cependant pas été accompagné de la création d'un nombre suffisant de postes permanents, ce qui fait peser, parfois lourdement, la gestion de ces plate-formes sur les chercheurs ou enseignants-chercheurs. De plus, l'absence de mécanismes permettant le transfert de connaissances entre les laboratoires de bioinformatique et les plate-formes, quand ils ne sont pas co-localisés, a éloigné certains chercheurs en bioinformatique des plate-formes et parfois aussi des expérimentateurs en biologie. Ainsi, l'Institut Français de Bioinformatique n'a été pourvu d'aucune mission en relation avec la recherche en bioinformatique. Pour remédier à cette situation il est essentiel de créer des mécanismes de rapprochement entre plate-formes et laboratoires de recherche en bioinformatique et modélisation. Cela permettrait la mutualisation des tâches de manutention et de développement de ressources, faciliterait le transfert de connaissances entre laboratoires et plate-formes et rapprocherait les modélisateurs et les bioinformaticiens des expérimentateurs.

Développement, visibilité et maintenance des ressources. Les nouvelles technologies produisent des volumes de données énormes qui posent de plus en plus de problèmes de stockage, d'intégration, de mise à jour et d'analyse. Il est devenu inefficace, voire impossible, de gérer ces données exclusivement au niveau des unités de recherche. Il faudra penser à une hiérarchie géographique de solutions en fonction des besoins. Une consultation nationale sur les moyens de calcul et de stockage en concertation avec l'Institut Français de Bioinformatique et le GIS France Grilles permettrait de dégager les besoins. L'Institut Français de Bioinformatique en coordination avec des projets internationaux devra jouer un rôle important dans la diffusion et la maintenance des logiciels, bases de données et ressources informatiques.

Logiciels. Il faut accroître la reconnaissance de l'activité de développement logiciel qui contribue à rendre plus visible la production scientifique. Promouvoir en France une meilleure culture du logiciel libre permettrait aussi de mieux s'inscrire dans l'environnement international, d'accéder à de nouveaux financements et d'obtenir plus de visibilité pour les développements méthodologiques. De façon générale, il existe un besoin d'appui informatique pour viabiliser les développements logiciels et promouvoir leur diffusion, voire leur industrialisation. Cela pourrait se faire sous la forme de mise à disposition d'ingénieurs en informatique pour des projets à durée déterminée.

Renforcer la capacité de la CID 51 à promouvoir l'interdisciplinarité. Actuellement les CIDs sont sous-utilisées car elles ne participent pas aux comités de visite des laboratoires et elles font très peu d'évaluations. Il est sans intérêt de multiplier les évaluations systématiques, mais des mécanismes devraient être créés pour solliciter les CIDs pour cer-

taines évaluations pointant des problèmes dans des équipes interdisciplinaires (en faisant attention à ne pas augmenter inutilement la charge liée à l'évaluation). La CID51 devrait pouvoir participer aux évaluations pour les promotions de niveau DR1 et supérieur. Enfin, une vraie articulation des activités des CIDs et de la mission pour l'interdisciplinarité permettrait une action plus efficace des deux structures.

Comité national de la recherche scientifique. « CID 51- Modélisation, et analyse des données et des systèmes biologiques : approches informatiques, mathématiques et physique ». *Rapport de conjoncture 2014*, [édition PDF en ligne]. ISBN : 978-2-271-08746-1. Disponible sur : http://rapports-du-comite-national.cnrs.fr/