43
Modélisation des systèmes biologiques, bioinformatique

La biologie à grande échelle est à l'origine d'une masse considérable de données qui concerne tous les niveaux du vivant :

- les gènes, les protéines et leurs interactions,
- les génomes, leur dynamique et leur évolution,
- les cellules, leur organisation et les mécanismes moléculaires sous-jacents,
- les organes et leur fonctionnement,
- les organismes et leur physiologie,
- les espèces et populations,
- les systèmes écologiques.

L'exploitation de ces données est au cœur de la CID43. Elle requiert à la fois des modèles mathématiques et physiques qui représentent les lois complexes du vivant, et des travaux en informatique pour simuler ou estimer ces modèles, fouiller les données, et pour intégrer toutes ces sources d'informations hétérogènes au sein de bases de données et de connaissances. L'objectif est une meilleure compréhension du vivant, avec des enjeux dans tous les domaines, médicaux, pharmaceutiques, environnementaux et agronomiques. Les années passées ont vu ces disciplines se développer de façon extraordinaire. Les articles les plus citées aujourd'hui, toutes sciences confondues, sont liés à l'exploitation informatique des données génomiques. Le mouvement continuera très certainement. La biologie de demain sera largement faite par des biologistes «secs», modélisateurs et/ou bioinformaticiens travaillant sur ordinateur plutôt qu'à la paillasse (dite «humide»). L'objectif de la CID43 est de favoriser les recherches dans ces domaines d'interface, en mettant en avant des chercheurs et des travaux innovants sur le plan méthodologique et répondant à des questions biologiques importantes. On trouvera dans la suite les principaux axes de recherche concernés, avec un regard plutôt biologique tout d'abord (quelles grandes questions biologiques ?), puis plutôt méthodologique (quels modèles ? quels algorithmes ? quelle intégration des données ?). Ces deux regards sont le plus souvent indissociables, mais ce mode de présentation facilitera la lecture par les tenants des différentes disciplines d'origine. Finalement, on tracera un rapide état des lieux, avant de conclure par les recommandations essentielles. Un glossaire explicitant les termes techniques est donné à la fin du document.

Génomique comparative et fonctionnelle

On a pu croire qu'après le séquençage du génome humain la course aux génomes allait ralentir. On assiste en réalité à une forte accélération. Cette accélération, facilitée par l'apparition de nouvelles techniques de séquençage rapides et peu coûteuses, est due à l'intérêt de comparer les génomes et d'explorer les divergences évolutives à

différentes échelles, depuis les études intra-spécifiques jusqu'aux analyses regroupant les grands domaines du vivant. A ce jour, environ 1000 génomes de bactéries, archées ou eucaryotes sont entièrement séquencés et disponibles publiquement. Grâce à ces données et aux nouvelles approches «phylogénomiques», c'est à dire se fondant non pas sur l'analyse de quelques gènes mais sur l'intégralité du génome, il devient envisageable d'élucider la phylogénie du vivant. Ce large échantillonnage taxonomique est particulièrement utile pour l'annotation des génomes (c'est-à-dire l'identification des gènes et autres éléments fonctionnels) par analyse comparative de séquences. Ces données offrent également une opportunité unique pour analyser la variabilité des répertoires de gènes et ainsi mieux comprendre l'adaptation des espèces à leur environnement (et notamment l'évolution de la virulence chez des organismes pathogènes). Le séquençage massif ouvre la voie à la «génomique des populations» – c'est-à-dire à l'analyse de la variabilité génétique au sein d'une population à l'échelle génomique. Ainsi, le projet de séquençage de 1000 génomes humains (dont la publication est prévue d'ici fin 2010), permettra non seulement de connaître l'histoire et la dynamique des populations de notre espèce mais également d'identifier les régions du génome soumises à pression de sélection (et donc impliquées dans l'adaptation de l'homme à son environnement), et d'analyser les mécanismes mutationnels à l'origine de la diversité génétique. Ces données serviront également de référence pour identifier des mutations impliquées dans des pathologies. Ainsi, le séquençage de génomes complets est une approche très prometteuse pour l'identification des mutations ponctuelles ou des aberrations chromosomiques impliquées dans des cancers. Enfin, le séquençage massif est aujourd'hui une technique de choix pour analyser la biodiversité au sein d'un écosystème (par exemple les projets de «métagénomes» pour analyser la diversité microbienne ou virale).

De plus, les progrès des techniques de séquençage ont ouvert la voie à de nombreux autres champs d'application. Désormais le séquençage «haut débit» permet d'analyser et quantifier l'expression des gènes (RNAseq), de détecter les interactions ADN-protéine (ChipSeq) ou ARN-protéine, de quantifier la méthylation de l'ADN (MethylC-Seq), d'analyser les processus de réplication, de recombinaison, l'organisation des chromosomes dans le noyau (C5), etc. Ainsi, il devient possible non seulement d'étudier la diversité génétique (entre individus ou entre espèces), mais également les modifications épigénétiques des chromosomes et leurs conséquences sur l'expression des gènes. Bref, ces nouvelles techniques révolutionnent l'étude de l'organisation et du fonctionnement des génomes.

Par ailleurs, les progrès techniques ne sont pas limités au séquençage. On assiste à la multiplicité des «–omics» (proteomics, metabolomics etc.). Toutes ces nouvelles approches ont transformé la biologie moléculaire moderne d'une science "pauvre en données" en une science "riche en données". Cette nouvelle donne représente un défi pour la bioinformatique. D'une part, le volume de données à traiter pose des difficultés algorithmiques importantes. D'autre part, les changements dans la nature même des données imposent de nouveaux développements (par exemple les méthodes statistiques développées pour l'analyse de données de puces à ADN ne sont pas directement transposables aux données de RNAseq ; les outils d'alignement de séquences couramment utilisés – tels que BLAST – ne sont pas adaptés aux nouveaux types d'analyse de séquence). Les enjeux sont considérables car ces nouvelles données permettent d'envisager, dès à présent, des analyses intégrées allant des séquences complètes des génomes aux conséquences phénotypiques de mutations, en passant par les aspects structuraux et fonctionnels sur les différents acteurs cellulaires. Face à ce volume croissant de données complexes et hétérogènes, l'intégration des données couplées à des analyses bioinformatiques comparatives et prédictives est cruciale pour réaliser la description étendue de la fonction d'un gène et de la compréhension de son rôle non seulement au niveau moléculaire, mais également aux niveaux supérieurs des complexes macromoléculaires, des voies cellulaires, de la cellule, de l'organe et de l'organisme.

Bioinformatique structurale

Le défi de la bioinformatique structurale est d'établir des liens entre la structure des macromolécules biologiques et leurs fonctions dans la cellule ou l'organisme. Sur le plan structural, les nombreux niveaux de complexité des molécules biologiques (ARN, ADN, protéines) doivent être pris en compte : structure primaire, secondaire, tridimensionnelle, quaternaire et structure des complexes multimoléculaires. Sur le plan fonctionnel, ce sont les aspects organisationnels de la matière biologique qui doivent être décryptés : aspects dynamiques et évolutifs, déplacements, assemblages, modes d'action, interactions, régulations et modulations des macromolécules et des systèmes macromoléculaires.

La bioinformatique structurale s'appuie d'une part sur les nombreuses données fournies par les grands projets de génomique et de génomique structurale, et d'autre part sur les données hétérogènes apportées par les diverses branches de la biologie, les domaines interdisciplinaires en émergence et les nouveaux développements technologiques. On peut définir quatre grands champs d'action pour la bioinformatique structurale :

• Un premier domaine d'action concerne le repliement des protéines et la prédiction de structures : identification et classification de motifs structuraux, développement de méthodologies comparatives au niveau structural, phylogénie structurale, analyse structurale prédictive des séquences/structures (ADN, ARN, protéines), problème inverse du repliement, modélisation par homologie à grande échelle.

• Un deuxième champ d'action relève de la modélisation et de la dynamique des macromolécules. Il se situe à l'interface avec les techniques de biologie structurale expérimentale : reconstruction de gros édifices 3D en utilisant des données hétérogènes (cryoEM, AFM, co-cristallographie, RMN du liquide et du solide, SAXS, méthodologies basées sur la fluorescence, imagerie moléculaire) ; ces secteurs nécessitent des couplages entre les outils informatiques actuels et les nouveaux développements méthodologiques, mais aussi la mise en

place de nouvelles démarches et méthodes d'analyse.

• Un troisième champ d'investigation concerne les machineries moléculaires avec la compréhension des mécanismes d'assemblage macromoléculaires (approches multi-échelles), des forces mises en jeu (expérimentations sur molécules uniques) et de la dynamique de ces assemblages. Les domaines concernés sont notamment les assemblages multiprotéiques, mais aussi l'auto-assemblage de membranes lipidiques, la simulation en dynamique moléculaire de gros systèmes associant protéines, membranes, ARN ou ADN (ribosome, facteurs de transcription, protéines membranaires modélisées dans leur environnement).

• Un quatrième champ d'action, concerne les interactions protéines-ligands ou substrats et les interactions protéine-protéine impliquées dans des voies métaboliques ou de signalisation cellulaire, ainsi que les mécanismes de modulation et de régulation de ces voies. Ce domaine d'étude s'appuie sur les données apportées par les techniques de la biologie moléculaire et cellulaire (telles que le double hybride, le TAP/MS, les puces à ADN et à protéines, l'imagerie des interactions in vivo de FRET ou par FCS). Il inclut la simulation des mouvements moléculaires (domaines, approches de ligands), les technologies d'ingénierie in silico de protéines ou de ligands, la prédiction des régions d'interaction et la prédiction des fonctions des protéines à partir de leurs réseaux de partenaires et de leurs régions d'interaction.

Imagerie *in vivo* des processus biologiques

Les progrès de la bioinformatique structurale, tels qu'ils viennent d'être rapportés se complètent aujourd'hui d'approches qui permettent de révéler et de quantifier la dynamique des processus moléculaires à l'échelle du vivant. Ces dernières années, les avancées en microscopie multidimensionnelle et multimodale couplées aux nouvelles techniques de marquage par sondes fluorescentes types GFP («Green Fluorescent Protein») ont révolutionné la biologie moléculaire et cellulaire. L'imagerie photonique dite à haute résolution spatiale et temporelle joue désormais un rôle essentiel pour sonder les processus moléculaires des interactions des protéines (expression et fonction des gènes) dans différents compartiments ou domaines cellulaires. Ainsi, des avancées considérables en biologie fondamentale ont déjà été obtenues dans la description des principales voies de transport membranaire, dans les mécanismes de tri et d'adressage des protéines et dans le maintien ou la défaillance de l'intégrité fonctionnelle de la cellule. Il est également établi que certaines protéines interagissent fonctionnellement et dynamiquement dans des sous-domaines de chaque compartiment, ceci pour assurer des fonctions vitales à des échelles spatiales et temporelles variées et sur l'ensemble du cycle cellulaire, que les cellules soient en division ou en interphase. Plus généralement le défi à relever désormais est d'acquérir une vue quantifiée plus complète de la physiologie de la cellule pour toutes les échelles d'observation spatiales, «nano», «micro» et temporelles. Enfin, une telle imagerie in vivo à une échelle d'introspection subcellulaire est aujourd'hui accessible dans des systèmes multi-cellulaires plus complexes.

Mais force est de constater que l'analyse des données de microscopie collectées, ne serait-ce que ces cinq dernières années, via différentes modalités de microscopie optique (modalités d'imagerie Nipkow Disk CLSM, FLIM-FRET, TIRF, SIM, STED…) et électronique, reste problématique. Il faut en effet traiter des quantités considérables d'images tridimensionnelles. Leur contenu est relativement nouveau et original, mais leur traitement est réalisé avec des algorithmes d'analyse d'images limités et peu adaptés désormais. Notons que l'analyse visuelle de ces données-images est devenue quasiment impossible puisqu'il s'agit d'inspecter des centaines de séquences temporelles d'images volumiques de plus en plus souvent acquises automatiquement (High Throughput imaging, High Content Screening), voire de les manipuler lors du passage d'une échelle à l'autre, par exemple dans le cadre d'imagerie corrélative (CLEM). Toutes les informations partielles ou descripteurs extraits automatiquement doivent être intégrés dans des modèles biologiques ou biophysiques à des fins de prédiction, pour des applications dans le domaine de la santé notamment. Dans un tel contexte, nous sommes confrontés à la fois aux problèmes de gestion de «masses de données», d'estimation en «grande dimension», de modélisation de «systèmes complexes» et de «changements d'échelles», dont il est aussi question dans le chapitre qui suit.

Biologie des systèmes

La biologie des systèmes étudie les processus biologiques considérés comme des systèmes complexes, multi-échelles et dynamiques. Elle apporte une vision intégrée de leur fonctionnement. Ces systèmes sont formés par des métabolites, des macromolécules, des cellules, des organes et/ou encore des organismes organisés en réseaux d'interactions.

Jusqu'à présent, deux approches majeures se sont développées parallèlement en biologie des systèmes. Elles se distinguent non seulement «historiquement» mais aussi par leurs objectifs, par la nature des données traitées et les méthodes mathématiques et informatiques auxquelles elles font appel. L'une, héritière de la biologie mathématique, modélise la dynamique de processus biologiques particuliers, à partir de données qualitatives et quantitatives généralement issues d'expériences à petite échelle. L'autre, plus récente, impulsée par le développement des «omiques» et de la bioinformatique, analyse et interprète des données produites par des expériences réalisées à haut-débit.

De manière intéressante et attendue, la convergence actuelle des deux sous-thématiques de la biologie des systèmes signe leur maturité respective. D'une part, les progrès en gestion, analyse et intégration des données «omiques» permettent leur utilisation en modélisation, pour l'amélioration des modèles et pour la validation des prédictions. D'autre part, la modélisation de grands réseaux est maintenant possible grâce aux avancées théoriques telles que la réduction de modèle ou les approches modulaires ou hiérarchiques. Les toutes

prochaines avancées de la biologie systémique, fruits de cette convergence, devraient faire définitivement admettre l'aspect biologique fondamental de la discipline.

En fait, les apports de la biologie systémique ces dernières années ont déjà modifié notre perception des processus biologiques et parfois au delà. Par exemple, l'importance des circuits et autres motifs (tels que les «feed-forward loops») dans les réseaux de régulation de gènes ouvre de nouvelles perspectives pour la compréhension des mécanismes moléculaires de régulation spatio-temporelle de l'expression génique. De nouveaux marqueurs/acteurs de pathologies humaines difficilement identifiables par les approches de gènes candidats classiques, ont été découverts suite à des analyses globales. Enfin, la robustesse des réseaux biologiques nous fait prendre conscience des limites de l'espace des systèmes biologiques sur lequel nous pouvons agir à des fins d'intervention, de contrôle ou de régulation. Seule une compréhension globale des réseaux d'interactions permettra de dégager de nouvelles pistes pour les développements en thérapie humaine. A ces fins, l'élucidation des liens entre génotype et phénotype par l'analyse des réseaux constitue un enjeu important de la biologie des systèmes actuelle, qui pourrait ouvrir de nouvelles voies, des biotechnologies à la médecine personnalisée.

Biologie intégrative

Les propriétés systémiques des organismes vivants tels que l'homéostasie et ses corollaires de robustesse et de résilience sont des propriétés émergentes à un niveau global qui doivent pouvoir être comprises à partir des dynamiques observées à un niveau «micro» (génétique, moléculaire, cellulaire) ou «méso» (tissu, organe). Il s'agit alors de concevoir les fondements théoriques et les stratégies expérimentales pour la reconstruction des dynamiques multi-échelles de l'organisme tout au long de son cycle de vie, corrélant génotype et phénotype et intégrant causalités ascendante (du micro vers le macro) et descendante (du macro vers le micro). C'est l'objet même de la biologie intégrative, qui fonde ses stratégies sur l'observation *in vivo* et l'intégration des processus biologiques à l'échelle cellulaire dans l'organisme entier.

Les initiatives des communautés internationales organisées autour du concept de «physiome» ou encore d'approches «systèmes complexes» se retrouvent autour d'une interdisciplinarité ou protocoles expérimentaux, acquisition de données et reconstruction phénoménologique puis théorique sont pensés ensemble. Il s'agit de fonder une description formelle des processus sur les mesures acquises à partir de l'observation *in vivo* des phénomènes aux échelles spatiales et temporelles adéquates. La reconstruction phénoménologique, étape d'extraction des mesures à partir des observations d'imagerie multimodale 3D+temps, nécessite des stratégies relevant des mathématiques appliquées et de l'informatique (filtrage, segmentation, détection et suivi spatio-temporel d'objets) pour conduire à un premier niveau de corrélations spatio-temporelles et de modélisation qualitative et prédictive. La description quantitative des phénomènes permet alors d'envisager une modélisation théorique et explicative des processus et leur simulation informatique (voir par exemple la morphogenèse de *Dictyostelium discoidum* en fonction de la production d'AMPc, ou la croissance racinaire chez *Arabidposis taliana* en fonction de la production et du transport de l'auxine, ou encore la formation des somites chez les vertébrés en fonction de l'oscillation couplée de l'expression génétique des facteurs de transcription her). Les prédictions fournies par les modèles font l'objet d'un retour vers l'expérience avec la mise en œuvre de stratégies de perturbation propres à la biologie (notamment : gain ou perte de fonction génétique, perturbation mécanique, ablation ou transplantation de cellules ou de tissus).

Le niveau d'organisation cellulaire est sans doute le niveau d'intégration obligé des dynamiques multi-échelles de l'organisme. En effet, la cellule est l'intégron des processus métaboliques, moléculaires et génétiques, et de ses interactions avec l'environnement cellulaire, multicellulaire et tissulaire. L'observation in vivo, la mesure et l'analyse des comportements cellulaires exprimés en termes de position, de trajectoire, de prolifération, d'interactions, de déformation, de mobilité, de différenciation et d'identité, font l'objet de nouveaux développement méthodologiques (voir par exemple les méthodes d'analyse cinématique de populations cellulaires en division dans un organisme). Ces mesures sont essentiellement obtenues à partir de l'analyse d'images 3D+temps in vivo, et la biologie intégrative motive actuellement nombre des développements dans ce domaine.

Le couplage entre les niveaux d'organisation du vivant nécessite l'observation et la mesure simultanée de paramètres à ces différents niveaux (par exemple observation et mesure simultanée de morphodynamiques cellulaires et d'expressions génétiques ou d'activités métaboliques au moyen de rapporteurs fluorescents). Le défi expérimental et méthodologique est considérable et peu de propositions ont été faîtes pour intégrer les dynamiques moléculaires, génétiques et cellulaires dans l'organisme. Les difficultés viennent notamment de l'absence de données adéquates en termes de résolution spatiale et temporelle pour modéliser la dynamique des réseaux d'interactions moléculaires et génétiques dans le contexte des déplacements et des divisions cellulaires dans l'organisme. Un autre pan de recherche en lien avec la biologie des systèmes est de construire des modèles spatialisés prenant en compte le devenir individuel de chaque cellule et de ses différents compartiments.

Enfin, la prise en compte de l'importance de la rétroaction des niveaux macroscopiques sur les niveaux microscopiques suscite de nouvelles voies de recherches tant au plan expérimental que formel. Il s'agit en particulier d'observer et de mesurer les coordinations à longue distance, mécaniques, chimiques, électriques dans l'organisme entier et de mettre en évidence les propriétés «immergentes» (du macro vers le méso et le micro) autant que les propriétés «émergentes» (du micro vers le méso et le macro). On peut signaler parmi les résultats marquants dans ce domaine, la mise en évidence des transductions mécaniques au cours de l'embryogenèse de la Drosophile.

Evolution et adaptation: du gène à l'écologie

L'évolution et l'adaptation forment un autre grand pan de la biologie. Les objets biologiques sont issus d'un processus d'héritage et de mutations, et ceci à toutes les échelles, du gène aux systèmes écologiques en passant par les espèces. Comprendre et retracer l'évolution de ces objets est souvent un pas décisif dans la compréhension de la fonction (par exemple, des gènes), dans l'élucidation de la structure (par exemple, des protéines), ou de la place dans un ensemble complexe (par exemple, des espèces au sein des écosystèmes). Retracer l'évolution apparaît également essentiel dans l'étude des maladies émergentes ou en évolution constante, telles que le SIDA, le SARS ou la grippe. Les études évolutives sont au centre des grands projets internationaux sur l'Arbre de la Vie, qui est la phylogénie de l'ensemble des espèces contemporaines et constituera un répertoire remarquable de la biodiversité globale.

Cette capacité à comprendre et modéliser le passé devrait prendre une nouvelle dimension dans ses applications écologiques avec les études sur le réchauffement climatique. Comment les espèces s'adapteront-elles ? Quels nouveaux équilibres entre espèces se formeront localement, quelles seront les impacts sur les sociétés humaines ? Ces questions se posent naturellement dans un contexte multidisciplinaire : la formalisation mathématique y est variée et ancienne; les méthodes informatiques avancées sont indispensables ; le couplage de logiciels climatiques et de dynamique des populations est à l'ordre du jour. La modélisation biologique et environnementale, ainsi que l'intégration des données génomiques, phénotypiques, écologiques, et climatiques, sont fondamentales et nécessaires.

Modélisation mathématique et méthodes statistiques

Toute science passe depuis des siècles par l'analyse des données et la modélisation, toutes deux de nature mathématique. La physique et la chimie sont sorties d'un mode descriptif au dix-huitième siècle avec la mise en équation de l'attraction des corps ou des réactions chimiques. Si la biologie a longtemps échappé à ces approches formelles – il y a des contre-exemples, tels les modèles proies-prédateurs ou les dynamiques de populations –, c'est clairement du fait de sa complexité et du manque de données quantitatives fiables.

Ce mot, "complexité", est celui qui caractérise en premier les domaines d'interdisciplinarité reconnus (le cerveau, l'univers, les sciences sociales). Suivant un mécanisme curieusement observé tout au long de l'histoire des sciences, la disponibilité des outils abstraits va de pair, et souvent précède leur emploi dans les disciplines "concrètes". Aujourd'hui les outils conceptuels, au premier rang desquels les outils mathématiques, ont acquis depuis environ un siècle, la capacité de traiter de tels problèmes complexes : les fonctions dérivables ont cédé le pas au mouvement brownien, les espaces euclidiens à des espaces de Hilbert de dimension infinie, les équations différentielles ordinaires partagent le terrain avec des systèmes dynamiques et des équations aux dérivées partielles, l'analyse des données ne se fait plus à la main mais avec des outils statistiques bien établis et toujours plus sophistiqués.

Cette complexité se reflète d'abord par une complexité accrue des modèles mathématiques, faisant souvent appel – ce qui est symptomatique de l'ampleur des problématiques – à des domaines relevant des maths dites "pures" : c'est le cas des systèmes dynamiques et des graphes, par exemple pour les processus biochimiques, leur fonctionnement et leur évolution. C'est également le cas de la théorie des jeux, pour ce qui concerne les systèmes écologiques et leurs fragiles équilibres. C'est aussi le cas de la géométrie appelée à jouer un rôle central, non seulement pour modéliser dans l'espace usuel les positions relatives des molécules pour mieux comprendre leurs interactions (par exemple, notion de site actif présenté à un substrat), mais surtout pour rendre compte dans des espaces de très haute dimension (espaces de lacets, par exemple) de la topologie des molécules biologiques (brins d'ADN, par exemple) et de leurs mouvements possibles (repliement, ouverture, …) sans doute le long de géodésiques dans ces espaces complexes. Cette complexité devrait aussi conduire à des modélisations multi-échelles pour lesquelles des évènements de natures différentes (évènements moléculaires de nature stochastique, évènements cellulaires et tissulaires plus déterministes) seront intégrés dans un même modèle utilisant des formalismes adéquats et différents selon les niveaux.

Mais les mathématiques plus traditionnellement tournées vers les applications ont connu, ces dix dernières années, un important développement vers l'interaction mathématiques-biologie. Les équations différentielles modélisent la dynamique moléculaire. Les équations aux dérivées partielles servent à comprendre toutes les échelles du vivant depuis le trafficking moléculaire dans la cellule, l'auto-organisation spatiale de communautés cellulaires, jusqu'à l'organe ou l'évolution darwinienne. Dans tous ces domaines la modélisation probabiliste et statistique est fondamentale, par exemple pour représenter les événements rares, ou les niveaux «individus centrés» plutôt que la population. Et bien sûr, tous les modèles devant être choisis sur des critères d'ajustement aux observations, tous leurs paramètres devant être estimés, toute hypothèse devant être testée, les statistiques sont centrales dans cette interdisciplinarité.

Il apparaît en effet clairement que le principal goulot d'étranglement dans les recherches actuelles et futures en biologie ne se situe pas dans la production de données, mais dans la capacité à les analyser et à les exploiter. Cette accumulation de données impose d'importants développements en statistiques. De nombreuses méthodes statistiques ont été publiées ces dernières années pour l'analyse de données de puces à ADN (étude des transcriptomes et des variants d'épissage, détection de variations de nombre de copie d'ADN, ou encore identification de sites de fixation de protéines sur le génome). Cependant, ces méthodes basées sur l'analyse d'un signal continu ne sont pas directement transposables

aux nouvelles approches par séquençage (RNAseq, ChipSeq, etc.) qui produisent des comptages. Ces développements requièrent une véritable pluridisciplinarité. En effet, le gain de puissance dans la détection de signaux que permettent les approches de séquençage «haut débit» s'accompagne également d'une plus forte sensibilité aux biais expérimentaux, liés à la préparation des échantillons biologiques ou aux techniques de séquençage elles-mêmes. Ainsi, le développement de ces méthodes statistiques requiert une bonne connaissance de la biologie et des techniques expérimentales pour pouvoir interpréter ces biais. Le séquençage 'haut débit' est cité ici en exemple, mais ce n'est évidemment pas le seul champ d'application des statistiques. En santé, les études d'association au niveau du génome et autres approches de «génétique génomique» requièrent le développement de meilleurs modèles (statistiques ou mécanistiques) reliant le phénotype au génotype et devront mieux tenir compte de la structuration des populations. Les études de génétique des populations et d'évolution, exigent également des développements méthodologiques importants.

Les enjeux du couple analyse statistique des données/ modélisation mathématique sont d'abord de comprendre des systèmes de plus en plus complexes et non-linéaires afin de donner du sens aux observations expérimentales caractérisées par des masses de données importantes et forcément incomplètes. Le développement de l'interface avec les sciences du vivant est très rapide avec de nombreuses nouvelles équipes investies. On peut penser que dans un futur proche, l'objectif sera aussi d'essayer de prédire (médecine personnalisée par exemple) et de contrôler ces systèmes (biologie synthétique par exemple), en particulier dans leurs fonctionnements pathologiques (réchauffement climatique par exemple).

Interface avec la physique

L'activité à cette interface est en plein essor. De nombreux biologistes se tournent vers les physiciens pour modéliser leurs systèmes, en même temps que de nombreux physiciens sont attirés par les enjeux de la biologie. Ces derniers amènent un bagage de savoir faire dans l'étude de «systèmes complexes» de toutes natures. En particulier, la physique théorique s'appuie sur la formalisation mathématique, l'utilisation assez systématique d'outils analytiques et de méthodes d'approximations, et le calcul numérique ou la simulation. Par la généralité de ces outils, les physiciens ont pu s'impliquer dans les sciences du vivant, dès lors que la modélisation apparaît possible et pertinente.

La question centrale des systèmes complexes peut se résumer ainsi : comment la connaissance des composants élémentaires ainsi que de leurs interactions permet-elle de comprendre et prédire le comportement collectif complexe émergent du système associé ? Les systèmes biologiques intègrent cette problématique : on connait assez bien les constituants cellulaires (ADN, ARN, protéines, lipides, métabolites etc.) dans une cellule vivante, et un peu moins bien leurs interactions, mais on est encore très loin de comprendre comment fonctionne une cellule. De même, on peut se demander comment les comportements d'individus simples conduisent à une société animale organisée (fourmis, abeilles, poissons). Et comment de tels systèmes s'adaptent-ils à leur environnement ? Le grand défi de la biologie est de comprendre la relation entre organisation des constituants «élémentaires» et les différentes fonctions biologiques, de l'échelle moléculaire à celle de l'écosystème, en passant par la cellule, l'organe et l'organisme. L'intégration de multiples échelles spatiales et temporelles est un point commun important entre la physique et la biologie. De ce fait, nous nous attendons à une implication croissante des physiciens en biologie intégrative.

Le rôle des échelles spatiales est un fil conducteur pour beaucoup des recherches à l'interface physique-biologie. Ainsi à l'échelle cellulaire et en dessous, l'agitation moléculaire peut rarement être ignorée ; la physique statistique (équilibre et hors équilibre) est alors un cadre approprié pour comprendre de nombreux processus moléculaires (cf. travaux sur molécules uniques) et cellulaires (du transport actif à l'expression génétique stochastique en passant par l'auto-assemblage de complexes supramoléculaires). La physico-chimie aussi joue un grand rôle à cette échelle. En allant vers les échelles spatiales plus grandes, pertinentes par exemple pour les tissus, on passe progressivement d'un domaine stochastique à un domaine relativement déterministe, rejoignant la physique des milieux continus, et en particulier des problématiques relevant de la mécanique et la biomécanique. Les travaux actuels de modélisation en biomécanique visent à décrire comment les objets biologiques s'organisent, fonctionnent et évoluent, en interaction avec leur environnement physiologique et physiopathologique. Pour décrire des processus comme l'angiogenèse, le remodelage osseux ou l'embryogenèse, il est indispensable de comprendre l'interaction entre les forces mécaniques développées par la cellule et la réorganisation de son cytosquelette et de ses points d'ancrage. Ces problématiques nécessitent des modèles couplant des champs physiques macroscopiques (écoulement de fluide, contrainte mécanique, diffusion de chaleur ou de matière…) à des systèmes biologiques complexes (cellules, tissus, organes). La mécano-transduction joue un rôle fondamental dans ces modèles. Elle permet à un système biologique de délivrer une information sur son environnement ou sur lui-même et de modifier ses propriétés sous l'action de contraintes physico-chimiques ou mécaniques diverses.

De façon générale, on peut dire que la modélisation, qui vise à donner un cadre synthétique et prédictif à des données ou des expériences, fait bien souvent appel aux nombreux outils et concepts de la physique, dont la physique statistique (à l'équilibre ou hors équilibre), la matière molle (polymères, gels, membranes, solutions), la physique non-linéaire, et la mécanique. Au-delà d'une description qualitative des interactions, la biologie va de plus en plus vers des données quantitatives, un cadre qui fait partie intégrante de la culture des physiciens. En retour, ces modèles sont susceptibles de proposer de nouvelles pistes expérimentales aux biologistes. Mais en biologie, les systèmes évoluent avec le temps et sont caractérisés par une grande variabilité. L'objet d'étude n'est donc jamais parfaitement défini, ni parfaitement reproductible. Par ailleurs, il existe peu de «lois» en biologie et beaucoup d'exceptions. Le physicien doit faire en sorte que les

cadres conceptuels qu'il définit prennent en compte cette diversité.

Une autre face majeure de l'interface physique/biologie concerne les techniques expérimentales et les technologies ou capteurs pour la biologie développés par les physiciens (voir par exemple la partie imagerie biologique ci-dessus). Ces techniques de pointe fournissent une masse de données quantitatives sans cesse croissante et le traitement de ces données exige des algorithmes particulièrement sophistiqués (segmentation d'image et suivi de traceurs étant deux exemples récurrents). L'interface physique/biologie se joue ici à la fois du coté «hardware» et du coté «software». La grande diversité des techniques d'acquisition de données fait que celles-ci relèvent généralement de commissions spécialisées. Mais l'aspect software au sens large, qui requiert à la fois des modèles et des algorithmes innovants, est très clairement inscrit dans la CID43.

Concepts et méthodes informatiques

Les progrès fulgurants des techniques d'acquisition expérimentale des données biologiques permettent d'accéder à des données de plus en plus massives, de plus en plus diverses, de plus en plus fiables, et de moins en moins coûteuses. L'avènement de ces nouvelles technologies devrait apporter aux biologistes des réponses à la fois plus précises et plus globales qu'auparavant. Cependant, fournir les méthodes permettant de traiter et analyser ces données pléthoriques pour aider à répondre aux questions biologiques est un véritable défi informatique.

Globalement, l'informatique doit fournir, en étroite collaboration avec les autres disciplines, les concepts nouveaux pour les approches multi-échelles, multi-niveaux, ou systémiques. Il s'agit d'étudier non plus les objets indépendamment les uns des autres, mais dans leurs interactions, et produire les méthodes pour aider à comprendre comment le comportement du système émerge de ces interactions.

La recherche en informatique pour la biologie est confrontée aux écueils liés à toute démarche de modélisation en sciences : la complexité des problèmes fait que tout modèle prenant en compte l'ensemble des paramètres connus sera pratiquement inutilisable. Du point de vue de l'informatique, cette notion se traduit d'abord en termes de complexité algorithmique. Les problèmes algorithmiques posés par la biologie sont, dans l'immense majorité des cas, NP-difficiles dès lors que la modélisation tient compte de tous les paramètres «réalistes» connus. Une difficulté majeure réside dans la recherche d'un bon compromis entre la précision de la modélisation et la complexité algorithmique de sa résolution. L'étude fine de la complexité du problème considéré en fonction des différents paramètres (complexité paramétrique) et/ou celle de son approximabilité, sont des préalables importants à la réalisation d'algorithmes ou d'heuristiques les plus précis et rapides possibles. Simplifier la modélisation initiale en relâchant ou omettant certains paramètres peut amener à un problème plus aisément soluble, typiquement polynomial. Cependant la simplification du modèle ne doit jamais perdre de vue l'impératif de réalisme biologique. Les problèmes théoriquement plus faciles, pour lesquels des algorithmes polynomiaux existent (la recherche de séquences dans les banques par exemple) motivent également des recherches poussées pour passer à l'échelle (parallélisme massif, programmation sur circuits programmables (FPGA) ou cartes graphiques (GPU), structures d'indexation sophistiquées, par exemple).

Une autre difficulté réside dans le caractère hétérogène, changeant, parfois peu fiable voire contradictoire selon les différentes sources, des données biologiques que l'informatique doit traiter. Les banques de données biologiques sont en constante évolution et il est impossible d'en contrôler la fiabilité. Les résultats expérimentaux sont toujours sujets à une marge d'erreur qu'il est parfois difficile d'évaluer. Les méthodes informatiques doivent impérativement tenir compte de ce fait, en étant d'une grande robustesse et adaptables à l'évolution des données.

Un axe essentiel de la recherche en informatique pour la biologie est l'algorithmique des séquences, des graphes et des structures discrètes en général. Ce champ de recherches fait face à de nouvelles problématiques liées au séquençage à haut-débit et au développement d'approches comparatives à très grande échelle pour l'étude du vivant à la lumière de l'évolution. Il est plus que jamais nécessaire de concevoir des algorithmes de traitement d'analyse non seulement des séquences, mais de données complexes qui soient efficaces, précis, robustes et passant à l'échelle des données massives. L'algorithmique des graphes est incontournable dès que l'on s'intéresse à la structuration de l'information génomique et aux relations entre les données biologiques. La biologie des systèmes, la biologie synthétique, la biologie structurale, la phylogénie notamment, sont des champs dans lesquels de véritables défis sont posés en termes d'algorithmique des graphes.

La classification et l'apprentissage automatiques constituent un autre domaine clé dont les applications sont nombreuses en bioinformatique (classification et annotation des transcrits issus d'expériences à haut-débit, classification de gènes selon leurs profils d'expression, apprentissage pour l'annotation automatique et pour la prédiction de structures moléculaires et de leurs interactions, par exemple). Ce domaine est en lien étroit avec la statistique et la modélisation probabiliste (modèles de Markov et Markov cachés, réseaux bayésiens, tests multiples, par exemple), et avec la théorie des langages pour certains aspects (grammaires stochastiques). Dans ce domaine il y a pléthore de données, ce qui est un avantage ; mais il y a aussi pléthore de paramètres, et souvent peu de données réellement fiables (c'est-à-dire avérées expérimentalement) sur lesquels les méthodes d'apprentissage peuvent se baser.

Plus généralement, le développement de méthodes pour enchaîner les étapes d'analyse à grande échelle des processus biologiques nécessite de continuer à développer des recherches dans d'autres domaines clés tels que la manipulation et l'intégration de données massives, complexes et hétérogènes, le calcul distribué et massif, l'analyse d'images et la géométrie computationnelle,

la modélisation formelle de systèmes dynamiques, la visualisation.

Finalement, une particularité de la recherche en informatique pour la biologie est qu'elle nécessite une réelle intrication de tous les domaines de la discipline, allant des plus théoriques, touchant les mathématiques, jusqu'aux domaines les plus proches du « hardware », en passant par la gestion des données et des connaissances ou l'algorithmique.

Etat des lieux

Une étude bibliométrique rapide[1] permet de positionner les recherches se faisant en France dans ces domaines. En matière de publications en bioinformatique[2], la France se place en 4ème position (~210 publications), très loin derrière les USA (~1660), mais aussi assez loin derrière l'Allemagne (~450) et l'Angleterre (~390), deux pays qui ont fortement investi le domaine depuis une bonne dizaine d'année, et qui bénéficient de la présence de laboratoires Européens. La France devance légèrement la Chine (~180, en progression rapide), le Japon (~150) et le Canada (~150). Pour ce qui concerne les publications en biologie des systèmes[3], la France (~60) est moins bien placée, loin derrière les USA (~520), l'Angleterre (~200), le Japon (~170) et l'Allemagne (~140), et à quasi-égalité avec l'Italie, l'Espagne, la Chine et le Canada (entre 60 et 70 tous les quatre). On peut voir dans ces chiffres la conséquence d'une certaine inertie ; la biologie des systèmes est depuis quelques années largement mise en avant au niveau mondial, mais la France n'a pour l'instant fait que peu d'efforts en termes de financement (un appel ANR plutôt restrictif depuis 2006, rien auparavant). Ce même facteur (avec un décalage d'une dizaine d'années ; aucun appel ANR spécifique) explique sans doute aussi les résultats seulement honorables en bioinformatique, par comparaison avec l'Allemagne par exemple.

Une autre mesure simple est la présence du thème bioinformatique dans les laboratoires du CNRS, telle qu'on peut la trouver dans l'annuaire des laboratoires sur le site du CNRS. En INSB 53 laboratoires sur 274 sont fléchés bioinformatique, en INEE 12 sur 80, tandis qu'on en trouve 5 sur 47 en INS2I, 1 en INC et aucun ailleurs. Même s'il ne s'agit pas d'une mesure réelle de l'activité, cela montre que les sciences de la vie au sens large ont bien compris l'intérêt de ces approches, mais que les mathématiques et l'informatique n'y ont encore mis que peu de forces. Il est également significatif que le thème biologie des systèmes (ou tout autre équivalent) n'apparaisse tout simplement pas sur le site. Enfin, on trouve difficilement 10 laboratoires (sur 1048) dont le nom évoque directement la bioinformatique ou la modélisation des systèmes biologiques, alors que de tels laboratoires existent en grand nombre à l'étranger. Ceci montre, si besoin, que l'effort vers l'interdisciplinarité que constitue la CID43 doit absolument se poursuivre et s'intensifier.

Sur la période 2003-2009 (7 ans donc), la CID43 (anciennement CID44) a assuré le recrutement (ou promotion CR1-DR2) de 39 chercheurs (+ 10 espérés en 2010). La pression était très forte, puisque nous avons auditionné près de 1000 candidats. Quelques autres recrutements sur les thèmes de la CID43 ont été faits dans d'autres sections (05, 07, 21, 22, 29 par exemple), mais avec généralement un caractère interdisciplinaire moins marqué. Notamment, ont été recrutés dans les sections de biologie des chercheurs appliquant des méthodes et programmes bioinformatiques, plutôt que contribuant à les faire progresser. Dans le même temps, de nombreux postes de bioinformatiques ont été affichés dans les Universités, pour répondre au besoin d'enseignements dans ces disciplines devenues indispensables à la biologie d'aujourd'hui. Ce bon niveau général de recrutement n'a malheureusement pas toujours été accompagné de la création de fortes équipes ou laboratoires, si bien que certains enseignant-chercheurs sont parfois isolés en ce qui concerne les aspects interdisciplinaires (ça n'est généralement pas le cas pour les recrutements CNRS et CID43, où cet aspect est pris en compte lors des concours).

Recommandations

On peut retenir des grands axes présentés ci-dessus quelques mots clefs : génomique comparative et fonctionnelle, biologie structurale, réseaux et systèmes biologiques, biologie intégrative, environnement et biodiversité, dont le développement dans les années à venir nécessitera à l'évidence des développements spécifiques en modélisation mathématique et physique, en statistique, en algorithmique, en imagerie, et en bases de données et de connaissances. Pour mener à bien ce programme, maintenir des recherches de premier plan en France, et développer harmonieusement cette interdisciplinarité au sein du CNRS, il importe de mettre en œuvre les recommandations suivantes :

Il faut accentuer les efforts en terme de postes interdisciplinaires, avec l'objectif de combiner au mieux : (1) réponses aux grandes questions de la biologie et au développement des approches à grande échelle ; (2) avancées des travaux méthodologiques, car ceux-ci accompagnent (voire précèdent) les progrès en biologie, et présentent souvent un intérêt propre dans leur discipline d'origine. Ces postes devraient principalement relever d'une section interdisciplinaire (de type CID43, pour assurer l'intérêt sur les deux versants), mais aussi des sections disciplinaires, et aller vers l'ensemble des instituts concernés (INSB, mais aussi INS2I, INSMI, INP, INEE…).

Il est indispensable que ces postes soient ouverts pour la plupart, et ne souffrent pas de fléchages ou coloriages trop contraignants. C'est particulièrement important dans l'interdisciplinarité où les viviers sont parfois restreints. Si le CNRS et les instituts souhaitent faire avancer une

1 Web of Science de l'ISI, période 2005-2009, nombre d'articles dont au moins un auteur est dans le pays considéré (les résultats sont très proches en considérant des périodes plus larges ou plus restreintes, sauf pour la Chine qui accélère nettement dans la période récente).

2 Publications dans la revue Bioinformatics (Oxford University Press) qui est la plus ancienne et a le facteur d'impact le plus élevé ; des résultats similaires sont obtenus avec d'autres revues comme BMC Bioinformatics ou PLOS Computational Biology.

3 Publications dans Biosystems, Molecular Biosystems, Systems Biology, Molecular Systems Biology.

politique scientifique particulière, il faut que les fléchages soient connus longtemps à l'avance, de manière à ce que les laboratoires aient le temps d'attirer des candidats de valeur.

Les recrutements doivent largement se faire au niveau CR2. L'interdisciplinarité de la CID43 est déjà bien établie, et on trouve d'excellents candidats, jeunes et ayant une réelle double compétence entre sciences formelles et biologie. Le niveau CR1 doit typiquement être ouvert et sans fléchage (cf. ci-dessus), pour recruter des chercheurs particulièrement brillants et ayant déjà un solide bagage, une vraie autonomie et une activité interdisciplinaire indiscutable.

Il faut repenser le suivi des chercheurs recrutés par la CID43 (le problème est sans doute analogue dans les autres CID). Ceux-ci sont à l'interface de plusieurs disciplines, mais l'évolution de leur carrière dépend de sections spécialisées qui leur préfèrent souvent des chercheurs davantage centrés sur le cœur de leur discipline. Il faut donc que la CID43 soit étroitement associée aux évaluations de ces chercheurs, en partenariat avec les sections dont ils dépendent directement.

Un objectif majeur est de développer les équipes ou laboratoires clairement situés à l'interface. Si ceux-ci commencent à voir le jour sur le versant biologique, ils sont forts rares sur le versant méthodologique (cf. l'état des lieux ci-dessus). A ce titre, une politique incitative doit être mise en place (relancée, car des efforts avaient été faits au tournant des années 2000), au travers de programmes CNRS, mais aussi de l'ANR qui s'est jusqu'à maintenant montrée peu interdisciplinaire. La gestion du suivi des chercheurs recrutés par la CID43 doit s'étendre à ces équipes et laboratoires interdisciplinaires.

Il faut développer l'activité de services en bioinformatique et en biostatistique, qui est indispensable aux biologistes à l'heure des données «haut-débit». Pour être performante, cette activité doit absolument être adossée à la recherche. En retour, la recherche bénéficie de services performants, par exemple lorsqu'il s'agit de récupérer des données ou de mesurer les progrès apportés par telle ou telle méthode. Une bonne part de l'interface entre biologistes et chercheurs en modélisation et bioinformatique, passe par les plateformes de services que ce soit pour l'acquisition et la gestion des données, ou les calculs de plus en plus lourds impliqués par le volume des données et la sophistication des méthodes. Le développement de cette activité implique essentiellement des recrutements d'ITA, qui stabiliseront et amplifieront les services aujourd'hui assurés par des CDD en nombre toujours croissant. L'adossement à la recherche et au développement permettra à ces ITA de rester performants et par conséquent d'accompagner la recherche en la faisant bénéficier des dernières avancées.

Glossaire

AFM : Atomic force microscopy
CLEM : Correlative Light Electron Microscopy
cryoEM : Electron cryomicroscopy
FCS : Fluorescence Correlation Spectroscopy
FPGA : Field Programmable Gate Array
FLIM : Fluorescence Lifetime Imaging microscopy
FRET : Förster Resonance Energy Transfer
GPU : Graphics processing unit
Nipkow Disk CLSM : Nipkow disk Confocal Laser Scanning Microscopy
TAP/MS : Tandem Affinity Purification/Mass Spectrometry
SAXS : Small Angle X-rays Scattering
SIM : Structured Illumination Microscopy
STED : Stimulated Emission Depletion
TIRF : Total Internal Reflection Fluorescence