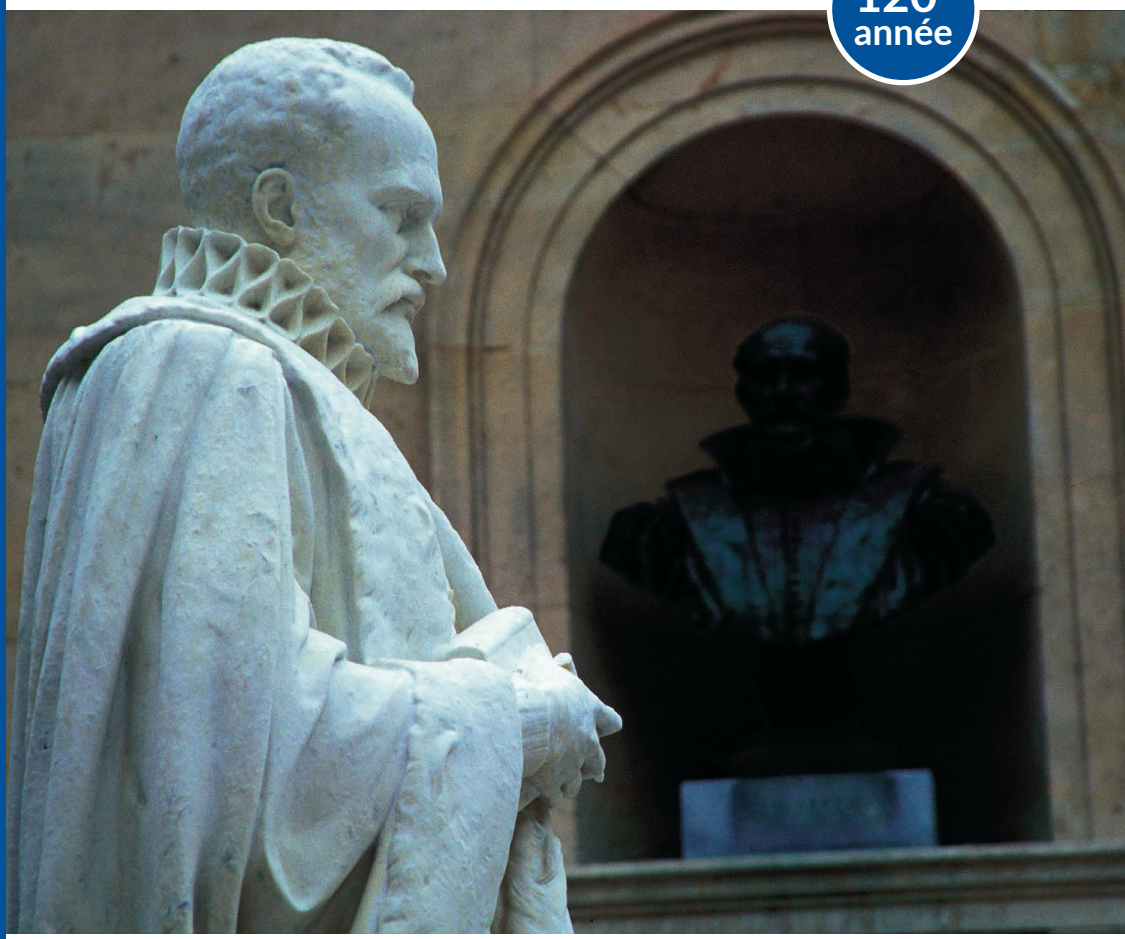
ANNUAIRE du **COLLÈGE DE FRANCE** 2019 - 2020

Résumé des cours et travaux

120e année



COLLÈGE DE FRANCE
1530

INFORMATIQUE ET SCIENCES NUMÉRIQUES*
(CHAIRE ANNUELLE 2019-2020)

Walter FONTANA
Professeur invité au Collège de France

Mots-clés : langage de modélisation, traitement moléculaire de l'information, transformation de graphes, causalité

La série de cours « La biologie de l'information – un dialogue entre l'informatique et la biologie » est disponible, en audio et vidéo, sur le site internet du Collège de France (https://www.college-de-france.fr/site/walter-fontana/course-2019-2020.htm), ainsi que la série de séminaires (https://www.college-de-france.fr/site/walter-fontana/p18441667353606966_content.htm). La leçon inaugurale « Le vivant et l'ordinateur : le défi d'une science de l'organisation » est aussi disponible (https://www.college-de-france.fr/site/walter-fontana/inaugural-lecture-2019-2020.htm). Celle-ci a également fait l'objet d'une publication : W. FONTANA, *Du calcul au vivant : le défi d'une science de l'organisation*, Paris, Collège de France/Fayard, coll. « Leçons inaugurales du Collège de France », vol. 291, 2019.

ENSEIGNEMENT

INTRODUCTION

La chaire Informatique et sciences numériques de 2019-2020 au Collège de France vise à mettre en exergue la biologie computationnelle. Le terme est souvent compris comme « bioinformatique » – une pratique du calcul au service de l'organisation, la recherche et l'analyse de grands ensembles de données en vue d'une compréhension

* Chaire créée en partenariat avec l'Inria.

W. FONTANA, « Informatique et sciences numériques », *Annuaire du Collège de France 2019-2020. Résumé des cours et travaux*, 120e année, Paris, Collège de France, 2023, p. 41-51 https://doi.org/10.4000/annuaire-cdf.18112.

prédictive. Un travail extraordinaire a été accompli dans ce domaine, et je crois qu'on y a accordé une attention suffisamment grande pour que l'organisation d'un autre cours se justifie (sans parler de mon manque d'expertise). Je voulais plutôt réinterpréter le sens de « calcul » en me concentrant sur la représentation des systèmes biomoléculaires (principalement constitués de protéines) comme un ensemble de règles « si-alors » dont les pré- et post-conditions capturent, à un niveau d'abstraction plus élevé, des résultats empiriques sur les mécanismes d'interaction. Par exemple, « si le domaine RGS d'Axin est lié à un domaine SAMP d'APC et GSK est lié à Axin et beta-catenin est lié à un domaine répété de vingt acides aminés d'APC, alors GSK phosphoryle beta-catenin. » Si cela n'a pas de sens, c'est parce que ce n'est pas le cas. C'est un fait empirique dépourvu de sens (autre que celui qu'il affirme), car il s'agit d'une seule brique Lego sans compagnons. Cependant, une fois associés à des dizaines ou des centaines d'autres faits de même nature, ces faits commencent à s'imbriquer dynamiquement et révèlent un système de comportement. Un tel style de modélisation considère un modèle comme un programme écrit dans un langage de programmation spécifique à un domaine. Les modèles de ce type peuvent être construits, débogués et analysés comme des programmes. Les fondements formels de la modélisation basés sur des règles reposent sur la transformation de graphes, car les pré- et post-conditions des règles sont exprimées sous forme de graphes de types spéciaux. Ainsi, sous cet angle, le terme « informatique » acquiert des connotations de la théorie du langage de programmation. Deux unités de cours sont donc consacrées à définir un langage appelé « Kappa », fondé sur des règles avec des exemples pour son application. Cependant, j'ai le sentiment qu'il doit y avoir une histoire plus large dans le contexte de laquelle la préoccupation pour ce type de modélisation est justifiée. J'ai choisi comme fil conducteur la notion quelque peu vague de « biologie de l'information », destinée à attirer l'attention sur le fait que l'information est toujours représentée physiquement d'une manière ou d'une autre, et que le traitement de l'information ne procède pas par manipulation d'une information éthérée et infiniment pliable, mais en agissant sur sa représentation physique d'une manière contrainte par cette représentation. Ainsi existe-t-il des conférences qui passent en revue des thèmes allant des cartographies génotype-phénotype à la précision de la réplication et de la reconnaissance moléculaire, et à l'apprentissage. Le choix, quelque peu éclectique, associe des sujets disparates de manière que les personnes habituées à apprendre des techniques détaillées pour s'attaquer à des problèmes ouverts dans des sujets bien fondés pourraient trouver frustrants. Enfin, la conférence inaugurale est une tentative de sonder en profondeur le terme « informatique » en mettant l'accent sur ses qualités générative et « chimique ».

Cours – Leçon inaugurale

La leçon inaugurale a un objectif ambitieux, distinct du cours lui-même : faire valoir que l'évolution et la dynamique de l'organisation fonctionnelle nécessitent un « mouvement » capable de construire son propre espace de phase, tout comme le développement biologique de l'organisme construit l'espace effectif (à ne pas confondre avec l'espace ambiant) dans lequel il se déploie. Je fais référence à ce mouvement comme à une « chimie », où le terme « chimie » doit être compris non pas comme une discipline, mais plutôt comme un concept. L'analogie qui me vient à l'esprit est celle de la Lune terrestre, qui n'est devenue lune que quand Galilée a découvert que Jupiter, aussi, avait des satellites.

Le cours s'est organisé autour de trois chimies et met l'accent sur leur idée commune : la transformation d'objets discrets structurés suivant des règles qui semblent jouir d'une existence autonome en faisant abstraction de la complexité sous-jacente dont ils découlent. L'archétype de cette idée est contenu dans la théorie des fonctions calculables, et plus précisément dans le *lambda*-calcul d'Alonzo Church. La première chimie que nous avons visitée est donc la chimie du calcul et les organisations algébriques dynamiques auxquelles le calcul donne lieu lorsqu'il est augmenté avec juste un tout petit peu de physique, comme la loi de l'action de masse. Le cas illustre comment une loi universelle (réduction *bêta*) peut se fractionner en de nombreuses règles qui peuvent être connues sans référence à la loi sous-jacente. L'idée d'une règle nous conduit alors à la seconde chimie, la chimie organique. La chimie organique repose sur la distinction entre règle et réaction. La règle ou schéma, spécifie uniquement les composants structurels nécessaires au déroulement d'une réaction avec les molécules auxquelles correspond le schéma. C'est le sens d'un « mécanisme » de réaction en chimie organique. Bien que les règles chimiques soient informées par la mécanique quantique, elles sont abstraites en ce qu'elles n'exposent pas explicitement leur origine. Parmi les organisations chimiques les plus intéressantes figurent les réseaux catalytiques et autocatalytiques qui constituent le fondement des systèmes vivants. Cela nous conduit alors à la troisième chimie, la chimie évolutive, celle des interactions entre protéines comprises comme agents ayant des domaines. Si l'idée de règle est analogue à la chimie organique, son contenu a évolué. À l'heure actuelle, on ne sait pas précisément quelle notion d'organisation fonctionnelle est à rechercher dans ce domaine. Une proposition est que l'organisation se compose d'une part de flux causaux câblés et, de l'autre, de flux causaux émergents et plastiques – communément appelés « voies » (de signalisation).

Dans l'ensemble, la leçon inaugurale n'est pas une vue d'ensemble du cours ; elle visait à présenter le défi de l'organisation fonctionnelle et une perspective pratique pour le relever : la modélisation fondée sur des règles de transformation de graphes, et les pistes théoriques qu'elle inspire.

Cours – La biologie de l'information : un dialogue entre l'informatique et la biologie

Cours 1 – L'encodage du phénotype et la topologie du possible

La première unité du cours portait sur le problème évolutif de l'innovation phénotypique, c'est-à-dire chercher à comprendre pourquoi les conséquences des mutations génétiques aléatoires ne sont pas aléatoires. Une réponse est que la modification d'un phénotype est rendue possible par le développement, c'est-à-dire qu'elle dépend de la cartographie du génotype au phénotype. Un modèle informatique de cette cartographie, à savoir le pliage des séquences d'ARN (« génotype ») en structures secondaires d'ARN à énergie libre minimale (« phénotype »), révèle quelques propriétés intéressantes.

Réseaux neutres

L'ensemble de séquences se pliant selon la même forme est mutationnellement connecté et étend un réseau à travers l'espace de séquence. Les réseaux neutres permettent aux populations de se diffuser à travers l'espace génotypique sans perdre leur phénotype actuel, tout en permettant l'exploration de nouveaux phénotypes

accessibles par mutation en périphérie du réseau. Les réseaux neutres illustrent comment la robustesse permet le changement. L'impact des réseaux neutres sur la dynamique évolutive apparaît dans les expériences informatiques comme des trajectoires évolutives avec des transitions soudaines.

Voisinage

La contiguïté entre réseaux neutres définit mathématiquement une notion de voisinage phénotypique qui n'est pas fondée sur la similitude et qui est suffisante pour définir la continuité et la discontinuité dans les trajectoires évolutives.

Couverture de l'espace des formes

Toutes les formes fréquemment réalisées se produisent dans un voisinage relativement petit d'une séquence aléatoire. Cela rappelle la propriété d'équipartition asymptotique, ou la notion d'ensemble typique, dans le théorème de Shannon sur le codage de source.

La plasticité reflète la variabilité

En généralisant la définition du phénotype pour inclure l'ensemble des structures alternatives (« états excités ») dans une bande d'énergie thermique à partir de l'énergie libre minimale (« état fondamental »), on observe que la variabilité génétique (l'ensemble des états fondamentaux réalisés dans le voisinage mutationnel d'une séquence donnée) est en corrélation avec la plasticité (l'ensemble des états excités réalisables par cette même séquence).

Cours 2 – La propagation de l'information génétique et phénotypique

La deuxième unité du cours était axée sur la propagation évolutionniste de l'information, comme vue à travers le modèle classique de Manfred Eigen, dans lequel les séquences se reproduisent avec erreurs dans le cadre d'un réacteur à flux. Le modèle donne naissance à un système d'équations différentielles décrivant l'abondance des séquences en fonction du temps, de l'attribution de la valeur adaptive (fitness) et du taux d'erreur. La non-linéarité légère peut être transformée et le système résolu avec l'algèbre linéaire standard. La solution stationnaire décrit une distribution de séquences, appelée « quasi-espèces ». Pour certains paysages de la valeur adaptive et modèles d'erreur, on constate qu'une reproduction précise détermine une longueur de séquence maximale au-delà de laquelle l'information représentée par la séquence la plus adaptée ne peut plus être préservée. L'acuité de ce seuil d'erreur dépend de la rugosité du paysage de la valeur adaptive et, pour certains paysages lisses, il n'y a pas de seuil net. Le modèle montre également que la quasi-espèce, comme un tout et non une séquence particulière, est l'unité de sélection. Ce qui compte, c'est la structure du paysage du fitness qui est peuplé par le nuage mutant entourant la séquence la plus apte ; un meilleur environnement peut amener une séquence moins adaptée à dominer une séquence plus adaptée. L'inclusion de réseaux neutres (cours 1) dans le modèle donne un seuil d'erreur phénotypique qui se produit à un taux d'erreur plus élevé que le seuil génotypique : la transmission du génotype peut être perdue sans que cela implique une perte de transmission du phénotype. Mathématiquement, le modèle des quasi-espèces est une version moléculaire du modèle de sélection par mutation de Crow-Kimura en génétique des populations. La différence est que, dans la version moléculaire, la

réplication et la mutation sont des processus se produisant conjointement, alors que, dans la version génétique de la population, ils se produisent indépendamment.

Cours 3 – La propagation de l'information moléculaire : petits circuits

Le troisième cours a donné un aperçu des aspects moléculaires de la transmission d'informations plus pertinents que ce qui se passe dans les cellules. Une erreur dans la reconnaissance moléculaire est inévitable en raison de différences finies et souvent petites dans les énergies de liaison entre les molécules concurrentes, qu'il s'agisse d'acides nucléiques accouplés à leurs compléments, de molécules de signalisation se liant à des récepteurs ou, généralement, de substrats qui se lient aux enzymes. Cependant, on peut réduire l'erreur, du moins arbitrairement, par une dépense d'énergie dans le processus classique de relecture. La relecture peut être fondée sur une discrimination cinétique dominée par des différences entre les barrières d'énergie à la liaison ou une discrimination énergétique dominée par des différences entre l'énergie libre des états liés. Au-delà du problème de précision, la conférence s'est concentrée sur les approches traditionnelles du raisonnement sur la prise de décision cellulaire. Ces approches réduisent considérablement la complexité combinatoire en utilisant un arsenal de schémas de petits réseaux de réactions chimiques, modelés avec des équations de cinétique, pour implémenter un répertoire de comportements de base. Un examen des schémas significatifs commence par établir un lien d'équilibre, dont le comportement de saturation est à la base de l'ordre de réaction accordable dans le modèle classique des interactions enzyme-substrat. La liaison simple a été généralisée à la liaison multivalente, qui, lorsqu'elle est séquentielle, conduit à un comportement de seuil. Un scénario dans lequel l'enzyme et le substrat doivent tous deux se lier à un échafaudage avant de pouvoir interagir conduit à une transmission du signal lorsque l'échafaudage est présent en concentration faible, mais porte à l'isolation des ligands quand la concentration de l'échafaudage est élevée. En général, les circuits fondés sur la liaison à l'équilibre peuvent implémenter une variété de propositions logiques. L'unité de construction la plus élémentaire au dehors d'équilibre est peut-être la boucle « faire/défaire » (Goldbeter-Koshland), qui, dépendant des conditions de fonctionnement, peut fournir un comportement marche/arrêt ultra-sensible ou une forme élémentaire d'adaptation précise. Combiner de telles boucles en série en une cascade peut produire une amplification et leur combinaison en parallèle conduit à un seuil. L'addition des retours conduit à une hystérésis et, donc, à de la mémoire. Pris ensemble, ces schémas et d'autres génèrent quelque chose qui ressemble à un « langage de programmation » de comportements primitifs qui peuvent être combinés et imbriqués dans des circuits de signalisation d'une grande complexité.

Cours 4 – Modélisation dans un cadre basé sur des règles (de transformation de graphes)

L'unité du cours introduit l'idée de description des systèmes moléculaires en termes de transformations de graphes ou règles. En ne rendant pas explicites les détails mécanistiques, les modèles classiques du cours 3 peuvent facilement être utilisés pour *imiter* beaucoup des comportements cellulaires. Cependant, les détails mécanistiques peuvent avoir des conséquences significatives et ne peuvent pas être simplement ajustés comme des constantes de taux. De plus, la complexité combinatoire en raison

de la modification post-traductionnelle et à la formation de complexes est omniprésente. Afin de pouvoir comprendre les aspects mécanistiques et combinatoires de la biologie moléculaire, un cadre récent, connu sous le nom de Kappa, propose de représenter les interactions protéine-protéine selon les lignes de la chimie symbolique, où les molécules sont des graphiques par-dessus les atomes typés. La chimie fait la distinction entre une règle (ou un schéma) et une réaction. Dans une réaction, les molécules sont complètement spécifiées, tandis qu'une règle ne rend explicites que les aspects des molécules qui sont nécessaires pour qu'une réaction se produise. Par conséquent, une règle représente la transformation d'un schéma de graphe. Si le schéma est contenu dans une combinaison moléculaire, la règle s'applique et génère une réaction. Dans Kappa, nous parlons des protéines comme des agents qui disposent de ports ou de sites porteurs d'un état (par exemple, la phosphorylation), et sur lesquels les agents interagissent comme spécifié par les règles. Les règles sont informées par des découvertes biophysiques et biochimiques, mais les expriment d'une manière transactionnelle abstraite. Pour qu'une règle induise une dynamique, elle doit être équipée d'une ou deux constantes de taux stochastiques, selon qu'elle s'applique dans un contexte intra- ou intermoléculaire. Ces constantes sont des taux de probabilité qui dépendent du volume de réaction et constituent des paramètres importants pour le simulateur Kappa. Le simulateur implémente une chaîne de Markov en temps continu d'applications de règles qui font évoluer le contenu d'un mélange moléculaire. En substance, l'approche fondée sur des règles considère un modèle comme un programme écrit en langage graphique qui définit une notion minimale de mécanisme local. En plus du simulateur, la plateforme Kappa implémente un analyseur statique, des outils d'analyse causale et une interface utilisateur graphique de base pour tous les principaux systèmes opérateurs.

Cours 5 – Détails et exemples de modélisation basée sur des règles

Le cinquième cours a examiné en détail plus formel, la syntaxe du langage Kappa. De nombreux exemples de règles sont donnés pour illustrer son expressivité. Deux aspects du design minimaliste de Kappa sont explicités.

L'interface d'un agent ne peut pas avoir deux sites portant le même nom

Cela reflète l'idée que deux domaines du même type (exemple : deux domaines SH2 dans un RasGAP) ne sont généralement pas situés dans le même contexte structurel et pourraient, par conséquent, différer dans leurs conditions d'interaction, ce qui justifie des noms distincts. Cette restriction est cruciale car elle rend la complexité d'identifier une correspondance d'un dessin Kappa dans un graphe hôte au plus quadratique de la taille du dessin.

Une règle Kappa est conçue pour être locale

L'intention est qu'il ne devrait jamais être nécessaire de vérifier en dehors du dessin pour déterminer si un dessin est contenu dans un graphe hôte. Cependant, si le dessin gauche d'une règle se compose de deux composants connectés, la règle peut induire un événement de réaction intra- ou intermoléculaire. La distinction est importante, car la constante de vitesse des deux cas diffère dans leur volume de dépendance. De telles règles obligent le simulateur à surveiller les informations globales, causant ainsi une augmentation substantielle du temps de calcul. Des modèles Kappa simples sont utilisés pour illustrer l'architecture d'un fichier

d'entrée. Un modèle plus avancé introduit des détails mécanistiques de base juste pour copier un polymère. Ce modèle met en œuvre une relecture basée sur la discrimination cinétique et énergétique comme expliquée dans le cours 3 et illustre comment l'erreur (entropie) peut conduire à une polymérisation énergétiquement défavorable. Bien que le modèle soit soluble analytiquement à l'état stationnaire, sa dynamique serait peu pratique à exprimer avec les équations habituelles de la cinétique chimique. La démonstration et l'exécution du modèle introduit l'interface utilisateur graphique de la plateforme Kappa.

Cours 6 – Causalité dans les modèles fondés sur des règles

En biologie des systèmes, les modèles basés sur la réaction sont souvent petits (comparé à la complexité du système), car ils résument drastiquement un mécanisme en conséquence d'une compréhension ou une hypothèse antérieure. Les modèles plus explicites basés sur des règles mécanistiques sont plus proches d'une représentation formelle de faits empiriques considérés comme des briques qui peuvent être assemblés pour construire des systèmes plus grands sans préjuger d'un comportement particulier (en fait sans une compréhension antérieure). Néanmoins, cela les rend aussi plus difficiles à comprendre, ce qui crée le besoin de concepts et d'outils de calcul pour mieux les analyser. Parmi les notions qui semblent particulièrement utiles pour comprendre les modèles basés sur des règles, il y a la causalité, comprise comme l'analyse rétrospective de la manière dont un événement ayant un intérêt s'est produit. L'unité 6 introduit des approches informatiques aux aspects de causalité dans le contexte fondé sur des règles :

(i) l'influence statique d'une règle sur une autre, définie en termes de chevauchements de dessins ;

(ii) l'influence dynamique entre les règles définit à quel point l'application d'une règle modifie la propension à l'application d'une autre. Une visualisation du réseau changeant d'influences dynamiques entre les règles est discutée dans le contexte d'un modèle d'horloge moléculaire (KaiABC) chez les cyanobactéries ;

(iii) une causalité de trace, définie comme une relation de précédence partiellement ordonnée représentant la mesure dans laquelle les événements, déclenchés par l'application de règles, auraient pu être permutés dans des histoires alternatives. Pour être utile, la causalité fondée sur la trace nécessite une notion de compression causale. En gros, la compression élimine les boucles causales dans lesquelles une série d'événements conduit à un état du système qui est équivalent à un état précédemment visité, en ce qui concerne la réalisation de l'événement d'intérêt ;

(iv) causalité contrefactuelle, dans laquelle l'événement X provoque l'événement Y si l'absence de X avait dû aboutir à l'absence de Y. Alors que la causalité de trace est « positive » en ce qu'elle enregistre les événements qui rapprochent un système d'un événement cible, la causalité contrefactuelle donne un aperçu des relations « négatives » ou inhibitrices entre les événements. Par exemple, dans un fait particulier, un événement X pourrait être la cause d'un événement Y, car X a empêché que l'événement Z se produise, ce qui, s'il s'était produit, aurait empêché Y.

Cours 7 – Complexité combinatoire dans les assemblages

La modélisation fondée sur les règles met en relief les aspects combinatoires de la signalisation. Mais dans quel sens ces aspects pourraient-ils être importants pour les

processus cellulaires ? Ce n'est pas évident, par exemple, que la combinaison exacte de résidus phosphorylés sur une protéine soit critique pour le traitement ou la propagation de l'information, ou si c'est juste la quantité globale de phosphorylation qui compte sur la base de la modification de la charge électrostatique globale de la protéine. Le niveau d'abstraction qui a permis la conception du langage Kappa n'est pas idéal pour exprimer l'état combinatoire au sein d'un agent (protéique) ; par contre, il est exceptionnellement bien adapté pour exprimer la combinatoire des interactions de liaison dans un complexe d'agents protéiques. C'est bien connu que les interactions protéine-protéine sous-jacentes à la signalisation cellulaire sont modérées par une variété de protéines d'échafaudage qui rassemblent les enzymes et leurs substrats. De récents efforts de modélisation suggèrent que l'assemblage d'un échafaudage pourrait être un processus où les éléments combinatoires peuvent avoir un impact substantiel qui peut être adéquatement représenté dans Kappa. L'aspect combinatoire intervient lorsque les protéines d'échafaudage peuvent se réunir. Un ensemble connecté de protéines réunies peut former un réseau ou une surface rassemblant de nombreuses instances d'une enzyme avec de nombreuses instances d'un substrat, ce qui augmente considérablement leur fréquence d'interaction locale et donc le taux catalytique. Dans le cas particulier où l'échafaudage consiste en des polymères linéaires, le taux catalytique en équilibre peut être calculé exactement en fonction d'affinité et d'abondance des monomères d'échafaudage pour le cas continu (concentrations) et le cas discret (nombres de particules). Toutes les interactions groupées sont soumises à un effet (« l'effet prozone »), dans lequel de faibles concentrations d'échafaudage favorisent les interactions entre les ligands, alors que des concentrations élevées les opposent en isolant les ligands sur différentes molécules d'échafaudage. L'avantage d'un système polymère est que le taux catalytique peut être facilement régulé et que la combinatoire atténue fortement l'effet prozone. En mettant en interaction des protéines d'origines distinctes, les agrégats d'échafaudage pourraient s'avérer essentiels pour combiner les informations issues de différentes voies de signalisation.

Cours 8 – Apprentissage cellulaire ?

Individus (et autres animaux) peuvent être considérés comme des modèles probabilistes de leur univers – un modèle incarné qui guide la perception et l'action. Beaucoup de travaux sur les réalisations biologiques de l'inférence probabiliste sont concentrés sur les circuits neuronaux, mais les organismes unicellulaires font aussi face à des environnements dynamiques et incertains qui exigent des comportements souvent formulés en termes cognitifs tels que la prise de décision et l'inférence. En effet, les idées issues des sciences cognitives ont été appliquées à la biologie des cellules individuelles et à celle des populations microbiennes. Pour un biologiste informaticien tourné vers la théorie, cela soulève la question de savoir comment un milieu moléculaire du type trouvé dans les systèmes de signalisation pourrait soutenir l'inférence probabiliste. Ceci est illustré par un exemple dans lequel un organisme unicellulaire, comme la levure, estime la dynamique de son environnement stochastique (celui qui balance entre le glucose et le galactose comme source de carbone) et utilise ces estimations pour réguler son métabolisme. Le modèle illustre comment les représentations pour soutenir l'inférence dans les modèles de Markov pourraient être incorporées dans des circuits cellulaires en combinant un schéma dépendant de la concentration pour coder les probabilités avec un mécanisme

moléculaire pour le comptage directionnel. Un cas différent, plus abstrait, vient d'Erik Winfree et de son groupe qui ont demandé quelles distributions de probabilités peuvent être réalisées en utilisant une version de réseaux de réactions chimiques dans lesquelles les réactions spécifient des interconversions entre des molécules dépourvues de structure (c'est-à-dire de types atomiques). Si les exemples peuvent être inventés et les stratégies pour les réaliser peuvent paraître irréalistes, ils donnent néanmoins le sentiment que la cinétique et la chimie de l'action de masse constituent un riche moyen d'inférence et que les cellules pourraient un jour être considérées comme des machines cognitives primitives.

Le reste de cette session a été consacré au résumé du cours.

Séminaires (en relation avec le sujet du cours)

La série de séminaires a permis de creuser profondément dans les sujets liés à la représentation, à l'origine et au traitement de l'information en biologie, sujets qui n'ont été que brièvement, ou pas du tout, abordés dans ce cours. À travers les dernières avancées en microscopie à feuille lumineuse, le dernier séminaire a donné un aperçu étonnant du monde moléculaire. Comme tel, il constituait un contrepoint délibéré au traitement principalement conceptuel et abstrait du cours.

Séminaire 1 – *The evolution of cellular individuality*

Le 8 novembre 2019, Eric Deeds (université de Californie, Los Angeles).

Séminaire 2 – *Graph rewriting and chemistry*

Le 15 novembre 2019, Daniel Merkle (université du Danemark du Sud, Danemark).

Séminaire 3 – *From molecules to systems: The problem of knowledge representation in molecular biology*

Le 22 novembre 2019, Jean Krivine (IRIF, Université de Paris).

Séminaire 4 – *Easy and hard in the origin of Life*

Le 29 novembre 2019, Eric Smith (Earth Life Sciences Institute, Tokyo).

Séminaire 5 – *Cells as cognitive creatures*

Le 13 décembre 2019, Yarden Katz (Harvard Medical School, Boston).

Séminaire 6 – *Thermodynamics of open chemical reaction Networks: Theory and applications*

Le 10 janvier 2019, Massimiliano Esposito (université du Luxembourg, Luxembourg).

Séminaire 7 – *Prediction in immune repertoires*

Le 17 janvier 2019, Aleksandra Walczak (ENS Paris).

Séminaire 8 – *Imaging sub-cellular dynamics from molecules to multicellular organisms*

Le 24 janvier 2019, Tommy Kirchhausen (Harvard Medical School, Boston).

RECHERCHE

La principale ligne de recherche poursuivie pendant mon séjour au Collège était axée sur l'utilisation de la plate-forme de calcul Kappa et des techniques analytiques pour comprendre l'importance de l'agrégation d'échafaudages de protéines dans les systèmes de signalisation cellulaire. Les échafaudages offrent des opportunités d'amarrage pour les protéines qui ne peuvent pas se lier directement. De cette manière, les échafaudages facilitent l'interaction. Le cas le plus simple est celui d'un seul échafaudage X qui peut se lier aux protéines E et S. Si la transmission d'un signal dépend de E (l'enzyme, par exemple) agissant sur S (le substrat, par exemple), alors la quantité du complexe EXS entièrement assemblé est cruciale pour le taux de transmission du signal. En équilibre, la quantité d'EXS en fonction de X à abondance fixe de ses ligands E et S est uni-modale : au début X augmente l'interaction entre E et S, mais lorsque X est en excès de ses ligands, il les isole de l'interaction car E et S sont liés à différentes instances de X. Cet effet dit « prozone » est bien connu.Des études empiriques et bioinformatiques suggèrent que dans certaines voies, les protéines d'échafaudage peuvent s'agglutiner en arrangements polymères. Nous avons donc entrepris une étude analytique computationnelle et approfondie d'un système d'assemblage en polymérisation, à la fois dans un contexte continu (où les concentrations réelles n'imposent aucune limite à la longueur maximale d'un polymère), et dans un cadre discret (où le nombre de particules détermine la longueur maximum atteignable de polymère). Les principaux résultats étaient les suivants :

(i) Dans le cas des polymères, l'effet prozone est « entropique », ce qui signifie qu'il résulte de la distribution de E et S entre les classes de longueur de polymère, car les classes de longueur les plus peuplées ne sont généralement pas assez peuplées pour isoler E de S au sein d'eux.

(ii) Un système de polymérisation fournit une augmentation significative de la concentration locale de E et S, agissant effectivement comme un petit compartiment qui augmente la fréquence de rencontre de E et S. Cela suppose que tout E peut agir sur n'importe quel S tant qu'ils restent liés au même assemblage d'agrégat.

Bien que le système de signalisation spécifique qui nous a inspiré ait une unité d'échafaudage (par exemple X) capable de polymériser, l'affinité entre les protomères X d'échafaudage semble être très faible, permettant principalement la formation de dimères. Cependant, il s'avère que ce même système contient également un deuxième type de protéine d'échafaudage (par exemple Y) qui a plusieurs sites de liaison d'affinité modérée pour X et peut ainsi agir comme une agrafe entre les unités X ou les X-oligomères. Dans les expériences de calcul, ces deux échafaudages subissent un processus d'agrégation d'équilibre en traversant un état métastable de longue durée à partir duquel une fluctuation nucléée déclenche un processus de croissance donnant lieu à un grand complexe fortement interconnecté d'unités X et Y. Ce phénomène repose sur une hypothèse (physique chimiquement justifiable) qui distingue les liaisons pouvant se former entre deux agents au sein d'un même complexe (fermant ainsi un cycle), et les liaisons entre agents appartenant à des complexes déconnectés. Dans ce dernier cas, l'énergie de liaison est diminuée par la perte d'entropie venant de la connexion de deux fragments indépendants ; alors que le premier cas n'entraîne pas un tel coût. Si le coût entropique est aussi bas que 9 kcal/mol à température ambiante, la conséquence est une préférence de 10^5 pour la

formation de liaisons intramoléculaires. L'analyse informatique de ces grands complexes révèle deux propriétés :

(i) Tous les types de liaisons sont toujours en équilibre pendant et après la croissance.

(ii) L'autocorrélation du plus grand complexe est élevée au regard des agents (étiquetés) qu'il contient, mais diminue rapidement en fonction de qui est connecté à qui. Ces propriétés indiquent qu'au sein du complexe, deux agents se dissocient vite, pour aussitôt, se réassocier avec d'autres agents au sein du même complexe. Ainsi, le voisinage d'un agent donné change rapidement et dynamiquement. Néanmoins, le complexe conserve son « identité » en termes des agents qu'il contient. Cela signifie que malgré les reconfigurations rapides des liaisons au sein d'un complexe, le complexe ne se désagrège jamais en fragments. Nous considérons cela comme une définition combinatoire d'une « phase liquide ». Nous croyons également comprendre la raison de la dynamique vitreuse de la condensation. Pour qu'un liquide se forme, il doit d'abord se produire une gouttelette, c'est-à-dire un complexe assez grand pour présenter une connectivité suffisante afin de survivre aux réarrangements de ses liaisons internes. Des expériences d'inoculation informatiques suggèrent que la taille des gouttelettes à elle seule n'est pas suffisante. La connectivité est aussi cruciale. Un complexe d'une taille donnée peut avoir de nombreuses configurations de liaisons (et donc un paysage énergétique complexe). À un extrême se trouve une configuration de liaison qui forme un arbre couvrant. Dans ce cas, toute dissociation de liaison fragmentera le complexe et entraînera sa disparition. À l'autre extrême se trouve une configuration qui réalise le nombre maximal de liaisons pouvant être atteintes de telle sorte que la longueur de tous les cycles soit maximisée. Dans cette configuration, le complexe peut tolérer une fraction substantielle de dissociations de liaisons sans se déconnecter ; en fait, il pourrait tolérer une fraction d'équilibre de liaisons dissociées. Le problème pour un complexe donné de taille appropriée est d'atteindre cette configuration de liaison interne (configuration d'énergie libre minimale). Ceci, croyons-nous, est à l'origine de la dynamique vitreuse et du phénomène de nucléation qui en résulte.

Dans des travaux indépendants, j'ai collaboré avec des collègues en informatique et en chimie sur des applications d'un cadre de transformation de graphe, analogue à Kappa, mais adapté à la chimie organique. Ces applications couvraient des domaines allant du suivi des atomes à travers des réseaux de réaction complexes à la conception de réseaux de réaction catalytique fondés sur des règles de la chimie des acides aminés dérivées de mécanismes enzymatiques.

PUBLICATIONS

FONTANA W., *Du calcul au vivant* : *le défi d'une science de l'organisation*, coll. « Leçons inaugurales du Collège de France », nº 291, Paris, Collège de France/Fayard, 2020.

ORTIZ-MUÑOZ A., MEDINA-ABARCA H.F. et FONTANA W., « Combinatorial protein-protein interactions on a polymerizing scaffold », *Proceedings of the National Academy of Sciences of the United States of America*, vol. 117, nº 6, 2020, p. 2930-2937, https://doi.org/10.1073/pnas.1912745117.