Modélisation mathématique de l'hématopoïèse et des hémopathies : développement, dynamique et traitement

Gurvan Hermange, Paul-Henry P.-H. Cournède, Isabelle Plo

Modélisation mathématique de l'hématopoïèse et des hémopathies : développement, dynamique et traitement

Auteurs:

Gurvan Hermange[1], Paul-Henry Cournède[1], Isabelle Plo[2,3,4]*.

1- Université Paris-Saclay, CentraleSupélec, Laboratory of Mathematics and Informatics (MICS), Gif-sur-Yvette, France.

2- INSERM U1287 (INSERM, Gustave Roussy, Université Paris-Saclay), Villejuif, France

3- Gustave Roussy, Villejuif, France

4- Université Paris-Saclay, Villejuif, France

* Correspondant: isabelle.plo@gustaveroussy.fr

Résumé

L'étude de l'hématopoïèse normale et pathologique, de par la complexité du sujet, requiert une approche multi-disciplinaire. L'utilisation de modèles mathématiques, lorsqu'il s'agit de comprendre la dynamique de populations de cellules, le développement de cancers ou l'effet d'un traitement, est particulièrement appropriée. Les modèles mathématiques, calibrés à partir d'observations expérimentales, peuvent être des outils d'aide à la décision en clinique, permettant par exemple de prédire l'effet d'un traitement ou d'en optimiser le dosage. Dans cette revue, nous commencerons par présenter différents modèles et formalismes mathématiques qui se sont développés au cours des décennies pour modéliser l'hématopoïèse, et qui sont encore pour certains à la base des travaux les plus récents. Nous aborderons ensuite les enjeux méthodologiques liés à l'inférence mathématique, permettant de s'assurer de la validité et robustesse des résultats. Enfin, nous terminerons par illustrer l'utilisation de modèles mathématiques dans trois champs d'application : l'initiation et le développement des hémopathies malignes, la dérégulation de leur hématopoïèse et leurs traitements.

Mots clés : Modélisation mathématique, analyse de données, cancers du sang, inférence statistique, traitement des hémopathies

Abstract

The study of normal and malignant hematopoiesis, due to its complexity, requires a multi-disciplinary approach. The use of mathematical models for understanding cell population dynamics, the development of cancers, or the effect of a treatment is particularly appropriate. Mathematical models, calibrated from experimental observations, can be used as clinical decision tools, allowing, for example, to predict a treatment's effect or optimize its dosage. In this review, we will start by presenting different mathematical models and formalisms that have been developed over the decades to model hematopoiesis, some of which are still the basis of the most recent work. We will then discuss methodological issues related to mathematical inference to ensure the validity and robustness of the results. Finally, we will illustrate the use of mathematical models in three fields of application: the initiation and development of hematological malignancies, the deregulation of their hematopoiesis, and their treatment.
Key words: Mathematical modeling, data analysis, blood cancers, statistical inference, treatment

Introduction

Les hémopathies sont des maladies qui induisent un dérèglement de l'hématopoïèse. Les comprendre – comprendre comment elles se développent, quelle est leur dynamique, comment les traiter, pour chaque individu – est essentiel pour une prise en charge optimale des patients. Cet objectif fait écho aux ambitions de développement d'une médecine personnalisée, c'est-à-dire adaptée aux spécificités de chacun, mais n'est pas encore atteint. Mais les progrès technologiques, permettant notamment de récolter des données hétérogènes, en grande quantité et à l'échelle de la cellule, concourent à sa réalisation.

Les modèles mathématiques, qu'ils soient déterministes ou stochastiques, constituent des outils pratiques et prometteurs pour l'étude du cancer – de leur initiation à leur traitement – ainsi que l'ont expliqué Altrock et al. [1]. Dans leur revue, les auteurs, au travers de nombreux exemples, illustrent les apports des modèles mathématiques dans la recherche contre le cancer. Pour reprendre leurs termes, « *les modèles mathématiques permettent de décrire un système par le biais de l'abstraction et du formalisme mathématique. Ils permettent d'extrapoler au-delà des situations initialement analysées, de faire des prédictions quantitatives, d'inférer des mécanismes, de montrer que des hypothèses biologiques sous-jacentes sont fausses et de décrire quantitativement les relations entre les différents composants d'un système* » [1]. Cela est particulièrement le cas lorsque les modèles sont confrontés à des observations expérimentales ; les données permettant alors la calibration des modèles, c'est-à-dire l'estimation quantitative des paramètres qu'ils mettent en jeu. Ces modèles, une fois calibrés, peuvent alors être utilisés comme outil de prévision et d'aide à la décision en clinique. Ceci nécessite, bien sûr, la construction d'un « bon » modèle (nous reviendrons plus tard sur ce que peut être la définition d'un bon modèle), mais également la mise en place de bonnes pratiques garantissant la robustesse des résultats, la validation des modèles, ainsi que la quantification des incertitudes.

Dans cette revue, nous commencerons par présenter un bref tour d'horizon des modèles mathématiques qui ont pu être construits pour décrire, que ce soit qualitativement ou quantitativement, l'hématopoïèse et les hémopathies. Cette section permettra de présenter la diversité des formalismes mathématiques existants, sans prétendre être exhaustive. Nous aborderons ensuite les enjeux méthodologiques qui surviennent lorsque l'on souhaite utiliser

les modèles pour l'analyse de données et obtenir des résultats quantitatifs. Cette section décrira quelques bonnes pratiques à mettre en place, et présentera pour les lecteurs biologistes ou cliniciens les problématiques fréquemment rencontrées par les mathématiciens lorsqu'ils se confrontent à l'étude quantitative d'un système biologique. Enfin, nous terminerons avec quelques exemples concrets d'application.

Tour d'horizon des modèles mathématiques

Les mathématiciens s'attellent à modéliser l'hématopoïèse depuis plus de cinquante ans. Dans sa revue, Pujo-Menjouet retrace une histoire de ces modèles, les différentes formes qu'ils ont pu prendre et les différentes tendances [2]. Les problématiques portant sur l'étude des oscillations périodiques survenant dans certaines maladies du sang ont rapidement intéressé les mathématiciens, dont les modèles permettaient de proposer des pistes quant aux causes potentielles des oscillations dans les quantités de cellules sanguines matures au cours du temps. En 1970, King-Smith et Morley suggéraient ainsi, à l'aide de simulations, que les oscillations survenant dans le cas des neutropénies cycliques pouvaient s'expliquer par la présence de mécanismes de régulation [3]. Plus tard, Mackey et al. formalisaient un modèle permettant de décrire les phénomènes d'oscillations dans certaines pathologies hématologiques, notamment les formes périodiques de la leucémie myéloïde chronique (LMC) [4]. Leur modèle repose sur une équation différentielle avec retard, c'est-à-dire une équation reliant les variations au cours du temps t d'une certaine variable – ici la quantité de globules blancs en circulation – à ses valeurs prises à un instant antérieur $t- \tau$ (et pas seulement à l'instant t, cas qui se modéliserait alors par une équation différentielle ordinaire). De nombreux auteurs ont depuis étudié l'hématopoïèse par des modèles d'équations différentielles avec retard [5]. L'introduction d'un retard dans les équations peut par exemple permettre de prendre en compte le délai nécessaire à la production de cellules matures à partir de cellules souches ou progénitrices [6] ou encore le temps nécessaire à la cellule pour effectuer son cycle [7]. Notons que l'introduction de ce retard n'est pas toujours un choix de modélisation, mais que les équations différentielles à retard peuvent résulter de l'intégration, suivant des variables de structure, d'équations aux dérivées partielles. Les équations aux dérivées partielles, liant certaines variables à leurs variations dans le temps mais aussi dans l'espace, permettent de prendre en compte des effets spatiaux, par exemple pour modéliser l'organisation d'îlots érythroblastiques [8]. Les modèles compartimentaux, dans lesquels certaines populations de cellules sont assignées à des compartiments (considérées alors homogènes) et où le passage d'un compartiment à un autre permet par exemple de modéliser les phénomènes de maturation / différenciation, sont largement utilisés pour modéliser l'hématopoïèse. Leur usage est particulièrement adapté lorsque l'on souhaite analyser des données expérimentales, les observations pouvant directement être associées à certains des compartiments permettant alors la calibration du modèle. Ils sont généralement formalisés par des équations différentielles ordinaires, avec parfois des non-linéarités lorsque l'on modélise des mécanismes de régulation, comme dans le cas de Marciniak-Czochra et al. [9]. Enfin, les mathématiciens se sont intéressés à la modélisation de la dynamique de l'ensemble des lignées hématopoïétiques, étudiées séparément (par exemple pour la lignée mégacaryocytaire [10], pour la lignée érythrocytaire [11], pour la lignée lymphocytaire [12]) ou ensemble [13].

Les différents formalismes cités jusqu'à présent (équations aux dérivées ordinaires, équations aux dérivées partielles, équations différentielles avec retard) sont déterministes. Étant données des conditions initiales ($t=0$) et des valeurs fixées (ou estimées) pour les différents paramètres, la sortie du modèle (par exemple la prédiction d'une certaine quantité de cellules

à un instant *t*) sera déterminée de façon unique, sans ambiguïté (Fig 1.C). À l'inverse, les modèles peuvent être stochastiques, c'est-à-dire introduire de l'aléatoire dans la description des dynamiques (Fig. 1.B). Alors que les modèles déterministes sont généralement valables lorsque l'on étudie la dynamique de populations cellulaires de grandes tailles, ils ne le sont plus lorsqu'il s'agit de s'intéresser à des comportements à l'échelle de la cellule, comme par exemple lorsque l'on souhaite modéliser le développement clonal d'un cancer à partir d'une unique cellule mutée (voir [14] pour une comparaison entre les approches déterministes et stochastiques dans le cas de modèles hématopoïétiques). Parmi les familles de processus stochastiques, mentionnons les processus Markoviens que l'on retrouve couramment dans la littérature. Il s'agit de processus dits « sans-mémoires », c'est-à-dire tels que leur état futur ne dépend que de leur état actuel (et non du passé). Ils ont été largement étudiés d'un point de vue théorique et sont relativement simples à simuler, ce qui justifie leur usage dans de nombreux modèles. Citons par exemple Catlin et al. qui proposent la calibration d'un modèle d'hématopoïèse à deux compartiments, reposant sur un processus de Markov caché [15]. Le terme « caché » signifie ici qu'il n'est pas possible d'observer la dynamique dans le compartiment des cellules souches ; Catlin et al. vont alors inférer ce qui s'y passe à partir d'observations au niveau des progéniteurs. Notons que la définition des processus Markoviens (sans-mémoire) n'est pas si restrictive, et que l'espace d'état peut être suffisamment grand pour modéliser des processus complexes par des processus de Markov. Néanmoins, ces processus ne seraient plus adaptés lorsque l'on cherche explicitement à modéliser des phénomènes de mémoire, par exemple une mémoire épigénétique [16]. Parmi les processus de Markov, mentionnons les processus de branchement qui permettent par exemple de décrire l'évolution au cours du temps d'une population de cellules – indépendantes les unes des autres – qui, lorsqu'elles se divisent, peuvent donner naissance à des cellules filles de différents types, suivant différentes probabilités (Fig 1.A). Un tel formalisme est par exemple utilisé par Xu et al. pour un modèle de division et différenciation de cellules hématopoïétiques, qu'ils calibrent à partir de données de *barcodes* cellulaires [17].

Mentionnons également deux processus Markoviens qui ont été couramment utilisés en génétique des populations [18], et dont on retrouve l'usage pour inférer le développement clonal dans certains cancers du sang (comme nous l'illustrerons plus loin) : le processus de Moran et le processus de Wright-Fisher. Il s'agit tous deux de processus à temps discret, c'est-à-dire que l'on s'intéresse à l'état du processus à différents instants $\{t, t+\Delta t, t+2\Delta t, ...\}$, par opposition aux modèles à temps continu. Contrairement aux processus de branchement pour lesquels le nombre de cellules total varie généralement au cours du temps, les processus de Moran ou de Wright-Fisher font l'hypothèse d'une population de taille constante.

Plus généralement, modéliser l'hématopoïèse, c'est définir un ensemble de règles et relations qui décrivent la dynamique et le comportement de différentes cellules hématopoïétiques, au cours du temps, éventuellement les unes par rapport aux autres, potentiellement dans l'espace. Les modèles n'aboutissent pas nécessairement à une formalisation sous la forme d'équations. Ainsi, les modèles multi-agents (un agent correspondant par exemple à une cellule), dont l'étude repose essentiellement sur des simulations numériques, se développent de plus en plus, favorisés par la diminution des coûts de calculs et l'usage de super-calculateurs. Bessonov et al. ont développé un logiciel basé sur un modèle multi-agents pour simuler et visualiser la dynamique de cellules hématopoïétiques [19]. Krinner et al. ont combiné un modèle déterministe avec un modèle multi-agents afin de décrire la dynamique de la granulopoïèse sous différents scénarios : une chimiothérapie ou une transplantation [20]. L'intérêt des modèles multi-agents repose notamment sur la possibilité de décrire des modèles aussi complexes qu'on le souhaite, mais ils s'accompagnent d'enjeux méthodologiques liés par exemple à leur calibration.

Enjeux méthodologiques et bonnes

pratiques

La modélisation mathématique d'un processus biologique – ici l'hématopoïèse – est une démarche qui nécessite une étroite collaboration entre mathématiciens et biologistes. Aboutir à un modèle (ou un ensemble de modèles potentiels) prend du temps : il s'agit de comprendre le phénomène étudié et de faire des choix selon la question de recherche d'intérêt et l'objectif poursuivi. Nous ne rentrerons pas ici dans une description méthodologique des étapes nécessaires pour aboutir à la formalisation d'un modèle, mais nous nous intéresserons aux étapes qui viennent ensuite, lorsque l'on souhaite calibrer le modèle à partir de données réelles, afin par exemple d'en interpréter les valeurs des paramètres, de faire de la prédiction, d'inférer des mécanismes biologiques non observés ou encore d'optimiser l'administration d'un traitement.

Nous considérons donc dans cette section avoir un modèle (ou un ensemble de plusieurs modèles potentiels) et que ce modèle est paramétrique, c'est-à-dire qu'il fait intervenir un nombre fini de paramètres inconnus qu'on souhaite estimer à partir d'observations expérimentales (notons que ce modèle peut également faire intervenir d'autres paramètres dont on connaît la valeur, qu'on appellera ici constantes). Ce modèle peut par exemple décrire la dynamique d'un système en simulant ses variables d'état $X(t)$ au cours du temps t : elles constitueront les sorties du modèle (qu'on nommera parfois « observations théoriques »). La figure 2 illustre, sur un exemple fictif, différents points abordés dans cette section.

Prétraitement des données

Pour calibrer un modèle, comme nous le verrons un peu plus loin, il faut pouvoir comparer la sortie du modèle à des observations qui lui correspondent. Les données brutes, issues d'expériences, sont rarement directement utilisables telles qu'elles. Un travail préliminaire sur les données (que nous appelons ici prétraitement) est souvent nécessaire. Chaque situation rencontrée est différente, mais évoquons cette problématique sur quelques exemples.

Tout d'abord, il est important d'évaluer l'incertitude que l'on a sur les observations. Les mesures ne sont jamais exactes : elles s'accompagnent d'un bruit de mesure (Fig 2.A). Pour les gérer, on définit généralement un modèle statistique pour les observations, le plus courant étant de considérer que les observations expérimentales sont distribuées suivant une loi gaussienne, de moyenne la valeur prédite par le modèle (correspondant à l'hypothèse choisie), et avec une certaine dispersion (variance de la loi). Mais d'autres modèles peuvent être plus pertinents suivant la situation rencontrée, par exemple si les observations sont nécessairement positives ou appartenant à un certain intervalle. Gérer les incertitudes, c'est aussi être en mesure de prendre en considération les événements non observés. Par exemple si l'on observe des divisions de cellules seulement jusqu'à un certain temps, il ne faut pas interdire dans le modèle que des cellules puissent se diviser après ce temps, sinon cela introduirait un biais dû à l'observation incomplète du processus.

Les données peuvent également ne pas être analysables directement sous leur forme brute. Par exemple, les données de séquençage du génome (*WGS – Whole Genome Sequencing*) de plusieurs cellules peuvent être structurées sous la forme d'arbres phylogénétiques, forme qui servira ensuite à leur analyse et à la calibration de modèles.

Enfin, une problématique fréquemment soulevée est celle de l'hétérogénéité, par exemple au niveau des cellules hématopoïétiques. À partir de l'expression de marqueurs de surface et de la définition de seuils, on peut trier les cellules en populations, par exemple cellules souches hématopoïétiques (CSH) ou progéniteurs multipotents (MPP) suivant l'expression du

marqueur $CD90^+$ chez l'homme tout en négligeant l'hétérogénéité entre cellules au sein des populations. Les modèles compartimentaux sont adaptés lorsque les données sont structurées sous cette forme. Or, comme des analyses *single-cell RNA-seq* ont pu le mettre en évidence récemment [21], les cellules hématopoïétiques constituent plus un continuum d'états qu'un ensemble de populations distinctes, ce qui pourrait se modéliser par des équations aux dérivées partielles. Ainsi, suivant la question de recherche à adresser, on peut choisir le niveau de détail à considérer dans les données et choisir le type de modèle en conséquence. La question de l'hétérogénéité se pose également lorsque l'on analyse les données de plusieurs individus. Peut-on regrouper les données ensemble, ou faut-il considérer les individus indépendants les uns des autres ? Des tests d'hypothèses (par exemple Mann-Whitney ou Kruskal-Wallis lorsque l'on a plus que deux échantillons, ou des hypothèses sur la nullité des variances dans le cadre des modèles à effets mixtes [22] peuvent être utiles pour indiquer si l'hypothèse d'observations distribuées suivant une distribution commune peut être rejetée, auquel cas il faudrait en toute rigueur traiter l'hétérogénéité inter-individuelle. Or, considérer les individus indépendants les uns des autres revient également à négliger un effet populationnel. Plutôt que d'un côté considérer les individus indépendants et d'un autre regrouper leurs données ensemble, on peut recommander l'usage de modèles à effets mixtes ou Bayésiens hiérarchiques qui prennent à la fois en compte la variabilité inter-individuelle et les similarités au sein de populations.

A priori biologiques

Les modèles mathématiques ne sont généralement pas créés *de novo*, mais reposent au contraire sur des *a priori* biologiques. C'est notamment un de leurs intérêts : ils incorporent au travers de nombreuses hypothèses des connaissances préalables sur le système biologique étudié. Les connaissances peuvent être précises, comme le choix de valeurs pour certaines constantes du modèle, ou plus vagues, comme par exemple les valeurs minimales ou maximales que peuvent prendre les paramètres du modèle qui seront à estimer. Le cadre Bayésien (dans lequel les paramètres du modèle sont considérés comme des variables aléatoires dont on souhaiterait estimer la distribution de probabilité à partir des observations à notre disposition) repose sur le choix de distributions *a priori* pour les paramètres. Avec une procédure d'estimation Bayésienne, les données viennent, en quelque sorte, actualiser notre connaissance sur le système étudié. Mentionnons Hoekstra et al. qui proposent l'utilisation de l'approche Bayésienne pour plus de transparence dans la présentation des résultats lors de la rédaction d'articles scientifiques [23]. Les *priors* sur les paramètres (ou les modèles, lorsque l'on en compare plusieurs entre eux) pourraient ainsi être définis de trois façons différentes, correspondant aux stéréotypes suivants : le sceptique, l'agnostique et le convaincu. Partant de ces trois différents *a priori*, on peut alors comparer les distributions *a posteriori* obtenues (c'est-à-dire après prise en compte des données) et voir si elles sont proches, ce qui suggérerait alors que les observations expérimentales sont suffisantes pour convaincre même les plus réticents.

Identifiabilité et sensibilité

S'assurer de l'identifiabilité d'un modèle est essentiel quand on souhaite pouvoir interpréter les valeurs estimées pour les paramètres. Un modèle est dit identifiable si on peut estimer sans ambiguïté la valeur de ses paramètres à partir de données expérimentales. Sinon, il ne l'est pas. La non identifiabilité est un problème car cela signifie par exemple que deux valeurs différentes pour un même paramètre pourraient conduire aux mêmes observations théoriques (Fig 2.B). Détecter les cas de non-identifiabilité n'est pas si simple, notamment lorsque le modèle est complexe et met en jeu de nombreux paramètres. Pour un exemple appliqué à l'hématopoïèse, citons Duchesne et al. qui évaluent l'identifiabilité de leur modèle d'érythropoïèse à partir d'une méthode basée sur la vraisemblance profilée [24]. Pour lever le

problème de non-identifiabilité, une fois celui-ci détecté, il faut soit avoir plus de données expérimentales, soit réduire le nombre de paramètres à estimer, en augmentant notre a priori sur les paramètres (par exemple en choisissant certains paramètres constants, fixés à des valeurs trouvées dans la littérature). Pour faire ce choix, une possibilité est de déterminer quels sont les paramètres qui ont le moins d'influence sur la sortie modèle : c'est-à-dire trouver quels paramètres peuvent varier sans fortement impacter la sortie du modèle. On dit alors que le modèle est peu sensible à ces paramètres ; l'approche consiste ainsi à mener une analyse de sensibilité. Parmi les différentes techniques existantes, mentionnons le calcul d'indices de Sobol.

Estimation des paramètres et de l'incertitude

Lorsque le modèle est formalisé, que les données ont été prétraitées et qu'on a défini les paramètres à estimer (après s'être assuré de l'identifiabilité du modèle), on peut procéder à la calibration du modèle. Cette étape revient à estimer les valeurs des paramètres telles que la sortie du modèle se rapproche le plus des observations. Cela nécessite de définir une distance entre les observations et la sortie du modèle, généralement déduite de la vraisemblance des paramètres étant donné les observations. La plus fréquemment rencontrée est l'erreur moyenne quadratique (*MSE - mean squared error*), associée à un modèle gaussien pour le modèle statistique des observations. En faisant varier les valeurs des paramètres, cette distance varie également, et on souhaite alors trouver le jeu de paramètres qui va la minimiser (Fig 2.C). Le problème ainsi défini est un problème d'optimisation pour lequel il est assez rare de trouver une solution explicite. On résout alors ce problème numériquement. Il existe de nombreux algorithmes pour cela, parmi lesquels on peut citer les algorithmes de quasi-Newton très efficaces mais seulement adaptés aux problèmes convexes [25] et la famille des algorithmes évolutionnaires dont fait partie l'algorithme CMA-ES (*Covariance Matrix Adaptation - Evolution Strategy*), développé assez récemment et qui montre de bonnes performances pour une large catégorie de problèmes, incluant ceux qui sont non-linéaires et en grande dimension [26].

L'approche présentée ci-dessus correspond à une approche dite fréquentiste dans laquelle on considère qu'il existe une vraie valeur pour les paramètres, qu'on souhaite estimer à partir des données. Dans la théorie Bayésienne, au contraire, les paramètres sont considérés comme des variables aléatoires dont on veut estimer la loi *a posteriori*, en combinant l'information *a priori* qu'on a sur elles (via la distribution *a priori*) et l'information provenant des données (via la vraisemblance, i.e. la probabilité d'avoir les données étant données les valeurs des paramètres). Avec une procédure d'estimation Bayésienne, on peut ainsi obtenir une estimation ponctuelle des paramètres (en choisissant par exemple la moyenne *a posteriori*), mais également un écart-type, ou mieux encore la probabilité que le paramètre soit compris entre telle et telle valeur. De nombreux algorithmes existent pour estimer la loi *a posteriori* suivant cette approche Bayésienne, dont le plus ancien est l'algorithme de Metropolis-Hasting. Cet algorithme repose sur la construction d'une chaîne de Markov (*MCMC – Markov Chain Monte-Carlo*), et a connu de nombreuses améliorations permettant une convergence plus rapide et en plus grande dimension (par l'utilisation d'algorithmes adaptatifs [27]). Citons également les méthodes ABC (*Approximate Bayesian Computation*) [28] qui connaissent un fort essor ces dernières années et qui sont notamment utilisées pour la calibration de modèles complexes, par exemple des modèles multi-agents pour lesquels il est facile de simuler le modèle mais compliqué d'exprimer une vraisemblance.

Une fois le modèle calibré, nous avons une estimation de la valeur de ses paramètres. Intuitivement, moins on a d'observations expérimentales et / ou plus elles sont bruitées, plus l'incertitude sur les paramètres (et, par propagation, sur la sortie) du modèle sera importante. Quantifier les incertitudes est primordial pour s'assurer de la confiance des prévisions du modèle. En théorie fréquentiste, plusieurs méthodes existent pour cela, dont par exemple les

techniques de *bootstrap* qui se basent sur du ré-échantillonnage de données. L'approche Bayésienne, quant à elle, permet naturellement de quantifier les incertitudes grâce à la distribution *a posteriori*.

Sélection et validation de modèles

Les modèles mathématiques permettent de décrire le comportement d'un système biologique à partir de différentes hypothèses. Différentes hypothèses conduisent ainsi à différents modèles. Sélectionner le meilleur modèle peut alors permettre de discriminer certaines hypothèses. Pour cela, plusieurs critères existent, par exemple le critère d'information d'Akaike (AIC) ou le critère d'information Bayésien (BIC) qu'on retrouve couramment. Sans rentrer dans les détails, il s'agit de faire un compromis entre la qualité de l'ajustement du modèle aux données (donc minimiser la distance entre les observations et la sortie du modèle, ou maximiser la vraisemblance) et le nombre de paramètres à estimer. Le principe (correspondant au rasoir d'Occam) est – lorsque l'on a plusieurs modèles – de sélectionner celui qui s'adapte le mieux aux données tout en étant parcimonieux. Les modèles avec trop de paramètres (trop de degrés de libertés) étant plus susceptibles de s'adapter aux données (on parle parfois d'*overfitting*), ils doivent être pénalisés.

Avec l'utilisation de ces critères, on peut alors sélectionner le meilleur modèle parmi ceux étudiés. Il reste alors la question de savoir si ce modèle est un *bon* modèle ; c'est-à-dire l'étape de validation des modèles. Mathématiquement, un bon modèle doit remplir certains pré-requis : être identifiable, parcimonieux, capable d'avoir généré les données, mais aussi de prédire des observations futures. Ainsi, une fois qu'un modèle est construit, calibré, ses capacités de prévisions doivent être évaluées à partir d'observations n'ayant pas été utilisées pour l'estimation des paramètres (par exemple des observations d'une cohorte de contrôle). Pour cela, on peut par exemple utiliser les observations d'un individu avant un temps T pour calibrer le modèle, puis utiliser les observations à $t > T$ pour mesurer l'erreur de prédiction (*mean squared prediction error*). Cette étape permet de s'assurer que le modèle est utilisable à des fins cliniques ; que ses prédictions ne seront pas erronées (Fig 2.D). Enfin, pour s'assurer qu'il décrit correctement le processus biologique modélisé, il restera à valider expérimentalement le modèle, en mettant au point par exemple de nouvelles expériences.

Exemples d'applications

Les modèles mathématiques peuvent trouver une application dans de nombreux domaines en hématologie. Nous choisissons dans la suite d'illustrer leur utilisation sur trois thèmes : apparition et développement des cancers, modélisation de l'hématopoïèse altérée et traitements. Nous ne prétendons pas faire une revue exhaustive de ces trois domaines d'application qui restent très vastes et nécessiteraient qu'un ouvrage leur soit consacré.

Inférer l'apparition et le développement des hémopathies malignes

Les hémopathies malignes sont encore trop souvent détectées tardivement, après apparition des symptômes et un développement clonal important. Comprendre leur dynamique d'apparition chez l'humain, à partir d'une cellule mutée, nécessite de pouvoir retracer le développement du cancer. Les modèles mathématiques sont particulièrement adaptés dans ce cas, avec pour objectif d'inférer ce qui n'a pas pu être observé (l'apparition et l'expansion clonale précoce), à partir de certaines règles décrivant le comportement des cellules.

Les modèles mathématiques peuvent par exemple être utilisés pour essayer de comprendre quelles sont les cellules à l'origine des cancers du sang. Modélisant la dynamique hématopoïétique par un processus de Moran, Traulsen et al. ont ainsi étudié le cas de l'hémoglobinurie paroxystique nocturne [29]. Avec une autre approche – basée sur des systèmes multi-agents et qui peut se généraliser à plusieurs cancers du sang – Haeno et al. ont étudié le cas des néoplasmes myéloprolifératifs (NMP) positifs pour la mutation *JAK2*^{V617F} [30]. Alors que l'hypothèse privilégiée consiste en un cancer de la cellule souche, Haeno et al. ont exploré des hypothèses alternatives, dont celle où la mutation *JAK2*^{V617F} surviendrait au niveau d'un progéniteur qui aurait acquis une capacité d'auto-renouvellement. Leurs conclusions nécessiteraient d'être confirmées par des expériences et des données, mais leur travail illustre la capacité des modèles mathématiques à tester différentes hypothèses biologiques et à en remettre en cause certaines. En l'occurrence, ils démontrent qu'il serait plus probable d'avoir l'acquisition de deux mutations – une mutation conférant une capacité d'auto-renouvellement puis la mutation *JAK2*^{V617F} – au niveau d'un ensemble de cellules se divisant fréquemment (les progéniteurs) plutôt qu'une seule mutation au niveau d'un ensemble de cellules souches ayant déjà une capacité d'auto-renouvellement mais se divisant peu fréquemment. Leur modèle, comme tout modèle, repose sur des hypothèses, et permet également d'identifier certains paramètres clés pour lesquels le modèle est sensible et donc sur lesquels des investigations devraient être menées : le nombre de cellules souches, leur fréquence de divisions, le nombre de divisions subies par les progéniteurs. La question du nombre de cellules souches contribuant à l'hématopoïèse est une question centrale. Lyne et al., à partir de simulations d'un modèle de Moran, montrent ainsi que, contrairement à l'intuition, une évolution clonale linéaire (généralement attribuée à un mécanisme de sélection naturelle) pourrait également être obtenue sans que les mutations n'aient d'avantage sélectif, à condition que le nombre de cellules souches contribuant à l'hématopoïèse soit faible [31]. Or, les estimations de leur nombre varient fortement suivant les auteurs. Dingli et al., modélisant la taille du compartiment des cellules souches hématopoïétiques actives N_{CSH} entre les différents mammifères par une loi allométrique $N_{CSH} \sim M^{3/4}$ (avec M la masse du mammifère), estiment qu'environ 400 CSH pourraient contribuer activement à l'hématopoïèse chez l'homme [32]. Plus récemment, Lee-Six et al. ont estimé que leur nombre se situerait plutôt autour de 100,000 [33], estimation qui semble aujourd'hui privilégiée notamment parce qu'elle se base sur des données expérimentales obtenues chez l'homme. Leur méthode repose sur le séquençage complet de colonies de progéniteurs hématopoïétiques d'un individu (sans pathologie connue), la construction d'un arbre phylogénétique à partir de l'information sur les mutations somatiques accumulées par les cellules, puis la calibration d'un modèle de Moran à partir d'une méthode ABC.

Les méthodes reposant sur la construction d'arbres phylogénétiques peuvent également être employées pour retracer l'histoire du développement clonal de certains cancers. Par l'identification d'ancêtres communs et l'hypothèse d'une accumulation des mutations somatiques linéaire au cours du temps [34] (qui en quelque sorte reproduirait une horloge moléculaire), on peut chercher à inférer l'âge d'apparition puis l'expansion d'un clone mutant responsable de la maladie ainsi que quantifier l'avantage prolifératif associé à la mutation. Ces méthodes ont notamment été utilisées dans le cas des NMP. Van Egeren et al., estimant les paramètres d'un modèle de Wright-Fisher à partir des données (structurées en arbres phylogénétiques) de deux patients atteints de NMP, ont ainsi montré que la mutation *JAK2*^{V617F} pouvait apparaître tôt au cours de la vie, lors de l'enfance ou l'adolescence, et qu'elle entraînait un avantage prolifératif au niveau des cellules souches [35]. Par une approche similaire, l'étude de 12 patients et le choix comme modèle d'un processus de naissance et de mort (qui est un processus de Markov), Williams et al. ont quant à eux montré que le taux de croissance des clones mutés *JAK2*^{V617F} était variable suivant les individus, qu'on pouvait le relier à la durée de latence entre l'acquisition de la mutation et le diagnostic de la maladie, et que des taux de croissance plus élevés étaient trouvés pour des clones abritant plusieurs mutations *driver* (motrices) [36]. Ils ont également trouvé que la mutation *JAK2*^{V617F} pouvait apparaître durant la vie fœtale.

Les méthodes reposant sur la construction d'arbres phylogénétiques ne sont pas les seules permettant d'inférer le développement clonal. A partir de l'information sur la fréquence allélique (*VAF – Variant Allele Frequency*) mesurée dans le sang périphérique pour près de 500,000 individus, et l'utilisation d'un modèle de branchement pour modéliser la dynamique des CSH, Watson et al. ont étudié le cas de l'hématopoïèse clonale et notamment quantifié l'avantage sélectif de plusieurs mutations, telles que $JAK2^{V617F}$, *SFRSF2* ou *DNMT3A* [37]. Notre équipe a également travaillé sur un modèle de branchement, calibré à partir de données de fractions clonales au niveau de progéniteurs pour des patients mutés $CALR^{m}$ ou $JAK2^{V617F}$, pour estimer des différences dans la dynamique d'apparition des mutations ($CALR^{m}$ apparaissant plus tardivement) puis l'avantage prolifératif entre ces deux mutations motrices des NMP ($CALR^{m}$ conférant un avantage prolifératif plus important au niveau souche). Nos travaux illustrent également l'utilisation de modèles mathématiques pour la mise en place de méthodes de dépistage précoces [51].

Au total l'utilisation de ces différents modèles ont permis des avancées importantes pour la médecine prédictive notamment en ouvrant la voie à la détection des patients à risque de développer des NMP (ou d'autres hémopathies malignes) et pointent sur l'utilisation de traitements visant à intercepter de façon précoce les clones malins (Fig 3.A).

Modéliser l'hématopoïèse altérée

De nombreux auteurs ont modélisé l'hématopoïèse comme un système dynamique. Mathématiquement, une hématopoïèse physiologique impliquerait que le système soit stable. Les maladies, cancers, ou stress hématopoïétiques peuvent alors être vus, dans ce cadre, comme une perturbation du système qui le déstabilise d'un état sain vers un état pathologique. C'est par exemple ce que font Stiehl et Marciniak-Czochra [38] où, partant d'un précédent modèle [9], étudient qualitativement les propriétés que devraient avoir les cellules leucémiques pour conduire à un développement de la maladie. Ils peuvent ainsi décliner leur modèle en plusieurs scénarios, suivant les valeurs prises par les paramètres, et par exemple décrire le cas des syndromes myélodysplasiques dans lesquels les cellules cancéreuses pourraient mettre plus de temps à se diviser mais avoir une augmentation de leur capacité d'auto-renouvellement, ou encore les lymphomes de Burkitt où la prolifération cellulaire serait plus importante chez les cellules mutantes comparées aux saines. La biologie des cellules cancéreuses est bien sûr plus complexe que celle décrite dans les modèles, mais ces derniers ont l'intérêt de se concentrer sur les paramètres les plus susceptibles d'induire la dynamique du cancer, et ainsi de mettre en évidence les caractéristiques des cellules mutées qu'on pourrait souhaiter infléchir, par l'usage de traitement, pour retrouver une hématopoïèse normale. On retrouve cette approche chez Moore et Li qui proposent un modèle de LMC (formalisé par un système d'équations différentielles non linéaires) dans lequel ils étudient l'influence des paramètres sur la valeur maximale atteinte par la concentration en cellules leucémiques [39]. Leur démarche correspond à une analyse de sensibilité ; au-delà de leurs conclusions biomédicales, ils illustrent une méthode qui permettrait de distinguer rigoureusement les paramètres d'un modèle qui devraient être estimés à partir d'observations de patients, de ceux qui pourraient être fixés constants à partir d'*a priori* biologiques (généralement issus de la littérature). En effet, les modèles mathématiques sont généralement construits pour être aussi simples que possible, avec un minimum de paramètres à estimer, sans quoi il devient difficile de tirer des conclusions de leurs analyses. On retrouve par exemple cette démarche de construction d'un modèle parcimonieux chez Andersen et al. [40], qui étudient par modélisation mathématique le lien potentiel entre inflammation et développement des NMP, notamment la progression entre thrombocytémie essentielle, polyglobulie de Vaquez et myélofibrose primaire. Notons que, dans les exemples précédents, on ne retrouve pas d'étape de calibration de modèles à partir de données longitudinales de patients avant traitement du fait de leur prise en charge thérapeutique. Ainsi, l'estimation des paramètres à partir de données de patients est le plus souvent effectuée

lorsque l'on étudie des modèles avec prise en compte d'un traitement, comme nous l'illustrerons dans la section suivante. Néanmoins, même si l'hématopoïèse murine diffère de celle chez l'homme, il peut être intéressant de calibrer des modèles d'hématopoïèse pathologique à partir d'observations longitudinales chez des souris. Ce type de données est par exemple utilisé par Bonnet et al. pour étudier l'hématopoïèse de stress [41]. Bonnet et al. modélisent l'érythropoïèse avec un modèle à six compartiments, des CSH jusqu'aux érythrocytes, en passant par différents types de progéniteurs. À un système d'équations différentielles ordinaires permettant de décrire l'hématopoïèse saine, ils ajoutent de la régulation, leur permettant alors de modéliser également l'hématopoïèse de stress, en particulier une anémie hémolytique associée à de l'inflammation. Ils induisent ce phénomène chez des souris par l'administration de phénylhydrazine et utilisent les mesures expérimentales pour calibrer leur modèle. Pour trouver les paramètres qui minimisent l'écart entre les observations et la sortie du modèle (et donc résoudre un problème d'optimisation), ils utilisent l'algorithme CMA-ES.

Comme nous l'avons montré à l'aide de quelques exemples choisis parmi une littérature très riche sur le sujet, modéliser l'hématopoïèse altérée – dans le cas des cancers du sang notamment – peut permettre de mettre en évidence les types de cellules mises en cause dans les dérèglements de l'hématopoïèse pour éventuellement identifier des pistes thérapeutiques (Fig 3.B). Nous présenterons dans le prochain paragraphe des exemples de modèles s'intéressant au traitement des hémopathies.

Comprendre, prévoir et optimiser l'effet d'un traitement

Pour reprendre les termes de Clapp et Levy [42], « *les modèles mathématiques sont un outil de recherche puissant qui peut être appliqué à la compréhension des leucémies et lymphomes. Ils peuvent identifier des mécanismes qui contrôlent la progression de la maladie, ou motiver et guider des expériences futures et des essais cliniques. En fin de compte, combiner modélisation mathématique, expériences et essais cliniques peut conduire à des améliorations significatives dans le traitement des leucémies et des lymphomes.* » Et de l'ensemble des cancers du sang, pourrions-nous ajouter. C'est lorsque les modèles sont confrontés à des données qu'ils nous semblent pleinement se révéler être de « *puissants outils de recherche* ». Michor et al. analysent ainsi à partir d'un modèle mathématique les données de 169 patients atteints de LMC et traités à l'imatinib (mesure du niveau de transcrit de fusion *BCR-ABL* dans le sang en cours de traitement) [43]. À l'aide d'un modèle à 4 compartiments décrivant la dynamique des cellules sous traitement, ils suggèrent que l'imatinib agirait sur les progéniteurs leucémiques mais pas sur les cellules souches, et prédisent alors que l'arrêt du traitement conduirait à une rechute. À la suite de l'article de Michor et al., l'étude sous un angle mathématique de la LMC et son traitement à l'imatinib a alors connu un important développement. Roeder et al. ont construit un modèle multi-agents qu'ils ont simulé et comparé aux données de deux cohortes de patients atteints de LMC et traités à l'imatinib [44]. Contrairement aux résultats précédents de Michor et al. [43], Roeder et al. montrent que les données peuvent également être en accord avec un effet du traitement au niveau des cellules souches, et prédisent une potentielle rémission à long-terme (comme ils le soulignent néanmoins, le développement de clones résistants au traitement pourrait réduire les chances de succès de la thérapie), rémission qui pourrait être accélérée en stimulant la prolifération des CSH. Foo et al. ont alors proposé par la suite un modèle prenant en compte les CSH quiescentes pour étudier l'effet combiné du G-CSF (*Granulocyte-Colony Stimulating Factor*) et de l'imatinib [45]. Ils prédisent, à partir de leur modèle et de données patients, que l'ajout de G-CSF pourrait augmenter le risque de résistance à l'imatinib et déconseillent cette option. L'interféron alpha (IFNα) pourrait être ce candidat qui, combiné à l'imatinib, améliorerait la réponse au traitement. C'est ce que suggère Glauche et al. [46] où ils étendent le modèle multi-agents de Roeder et al. [44] pour y inclure un effet de l'IFNα. Plus récemment, Bunimovich-Mendrazisky et al. [47], à partir d'un modèle de LMC faisant intervenir

des populations de cellules souches leucémiques, cellules matures leucémiques et lymphocytes T cytotoxiques, ont simulé l'effet de l'IFNα combiné avec l'imatinib. Dans leur article reposant sur la simulation de leur modèle sous différentes hypothèses, ils montrent que plusieurs scénarios seraient favorables à l'utilisation de l'IFNα qui, combiné à l'imatinib, pourrait permettre d'induire une réponse moléculaire complète et de l'accélérer. À la fois pour Glauche *et al.* et Bunimovich-Mendrazitsky et *al.* [46,47], la portée de leurs résultats est limitée par le fait que leurs modèles ne sont pas calibrés à partir de données de patients, mais ils proposent néanmoins des pistes intéressantes de thérapie qui pourraient être ensuite testées cliniquement. Comme Clapp et Levy [42] le soulignent dans leur revue, les modèles mathématiques peuvent en effet être utilisés pour tester différents scénarios *in silico*, et éventuellement orienter certaines pistes de recherche : chez Glauche *et al.* et Bunimovich-Mendrazitsky et *al.,* la modélisation mathématique peut être alors vue comme un outil permettant de sélectionner des pistes de recherche prometteuses ou pour construire des essais cliniques.

Nous avons mentionné plus haut l'IFNα comme traitement potentiel associé à l'imatinib dans le cas des LMC. Son utilisation dans un autre type de cancers du sang – les NMP non *BCR-ABL* – a également été démontrée puis étudiée mathématiquement par différents groupes. À partir des données d'une cohorte danoise (*DALIAH trial*, mesures de la VAF pour $JAK2^{V617F}$ au niveau des cellules matures), Ottesen, Pedersen et al. basant leur travail sur le modèle Cancitis [40,48], ont modélisé l'effet de l'IFNα sur le taux de mortalité des cellules souches mutées. Les dynamiques observées pour les patients de leur cohorte sont en accord avec leur modèle. De plus, ils démontrent l'utilité de leur modèle comme outil d'aide à la décision pour prédire le résultat du traitement à l'échelle du patient. Notre équipe a également étudié l'action de l'IFNα sur les progéniteurs hématopoïétiques en mesurant régulièrement l'architecture clonale des mutations chez des patients atteints de NMP, puis à partir de ces données, calibré un modèle mathématique permettant d'inférer l'action de l'IFNα au niveau des cellules souches initiatrices de la maladie, en distinguant l'effet suivant les clones hétérozygotes et homozygotes [49]. Étendant ce modèle pour prendre en compte les variations de posologies, Hermange et al. ont récemment suggéré que l'usage de doses d'IFNα suffisamment élevées seraient nécessaires pour induire une rémission sur le long terme dans le cas de patients présentant la mutation (que ce soit hétérozygote ou homozygote) $JAK2^{V617F}$ [50]. Pour calibrer ce modèle, notre équipe a utilisé une méthode d'estimation Bayésienne hiérarchique permettant à la fois d'estimer les paramètres à l'échelle individuelle (pour chacun des patients) ainsi qu'un effet populationnel.

Les quelques exemples précédents, malgré leur choix subjectif et ici axé sur les NMP, permettent d'illustrer l'usage de modèles mathématiques pour étudier l'effet de traitements, à partir de données patients ou *in silico*, pour ensuite faire de la prédiction ou de l'optimisation de traitement (Fig 3.C). Bien que nous n'ayons pas mis l'accent sur les questions techniques dans le paragraphe ci-dessus, les auteurs s'emploient également à mettre en place une méthodologie s'assurant de la robustesse des résultats, de leur reproductibilité et de leur validité. Soulignons néanmoins que les modèles mathématiques sont à chaque fois construits suivant certaines hypothèses. La validité des résultats est nécessairement conditionnée aux différentes hypothèses explicitées par leurs auteurs. Différentes hypothèses peuvent conduire à des interprétations qui diffèrent, comme sur l'exemple de Roeder et al. [44] qui obtenaient des conclusions différentes de celles de Michor et al. [43]. Les modèles mathématiques, lorsqu'ils sont utilisés pour élucider un mécanisme d'action d'un médicament par exemple, peuvent permettre de conclure en faveur ou en défaveur de certaines hypothèses ; ils ne remplacent pas l'étape de validation biologique mais la complètent.

Perspectives

Les hémopathies malignes sont des maladies complexes qui altèrent un processus biologique – l'hématopoïèse – dont il reste encore beaucoup à découvrir. Une approche multidisciplinaire, incluant une étroite collaboration entre mathématiciens et biologistes, est indispensable pour tirer profit du nombre croissant de données disponibles, et de leur complexité. Les modèles mathématiques, par une construction reposant sur des *a priori* biologiques – par exemples issus de données précédemment collectées et analysées – sont alors des outils permettant d'inférer des mécanismes biologiques, et dont on peut tirer profit à des fins de prédiction et d'optimisation de traitement. Nous avons souhaité présenter ici le travail méthodologique important qui accompagne la création de tels modèles, nécessaire pour s'assurer de la robustesse des résultats. La réussite des projets interdisciplinaires entre mathématiciens et biologistes repose avant tout sur une étroite collaboration et une volonté réciproque de partage de connaissances : les mathématiciens doivent faire l'effort de comprendre la biologie et les biologistes doivent accepter de se plonger (un peu) dans les équations. Les succès de ces collaborations, tels qu'illustrés par les exemples que nous avons présentés, permettent pas à pas de répondre aux ambitions d'une médecine personnalisée. Dans cette optique, les méthodes mathématiques d'apprentissage statistique *(machine learning & artificial intelligence*), adaptées pour traiter des données en grande dimension telles que des données *single-cell RNA-seq* par exemple, sont également très prometteuses et connaissent un fort essor en hématologie ces derniers années, approfondissant notre compréhension de l'hématopoïèse [21].

Remerciements

Nous remercions le Dr. Hana Raslova, membre du comité scientifique de la Société Française d'Hématologie pour la relecture de cette revue.

Références

1. Altrock, P. M., Liu, L. L. & Michor, F. The mathematics of cancer: integrating quantitative models. *Nat. Rev. Cancer* **15**, 730–745 (2015).
2. Pujo-Menjouet, L. Blood Cell Dynamics: Half of a Century of Modelling. *Math. Model. Nat. Phenom.* **11**, 92–115 (2016).
3. King-Smith, E. A. & Morley, A. Computer Simulation of Granulopoiesis: Normal and Impaired Granulopoiesis. *Blood* **36**, 254–262 (1970).
4. Mackey, M. C. & Glass, L. Oscillation and Chaos in Physiological Control Systems. *Science* **197**, 287–289 (1977).
5. Crauste, F. Delay Model of Hematopoietic Stem Cell Dynamics: Asymptotic Stability and Stability Switch. *Math. Model. Nat. Phenom.* **4**, 28–47 (2009).
6. Bélair, J., Mackey, M. C. & Mahaffy, J. M. Age-structured and two-delay models for erythropoiesis. *Math. Biosci.* **128**, 317–346 (1995).
7. Adimy, M., Chekroun, A., Touaoula, T. & Yang, K. Age-structured and delay differential-difference model of hematopoietic stem cell dynamics. *Discrete Contin. Dyn. Syst.* **Series B, 20**, 2765–2791 (2015).
8. Eymard, N., Bessonov, N., Gandrillon, O., Koury, M. J. & Volpert, V. The role of spatial organization of cells in erythropoiesis. *J. Math. Biol.* **70**, 71–97 (2015).

9. Marciniak-Czochra, A., Stiehl, T., Ho, A., Jäger, W. & Wagner, W. Modeling of asymmetric cell division in hematopoietic stem cells|regulation of selfrenewal is essential for effcient repopulation. *Stem Cells Dev.* **18**, 377–386 (2009).
10. Boullu, L., Pujo-Menjouet, L. & Wu, J. A Model for Megakaryopoiesis with State-Dependent Delay. *SIAM J. Appl. Math.* **79**, 1218–1243 (2019).
11. Crauste, F., Pujo-Menjouet, L., Génieys, S., Molina, C. & Gandrillon, O. Adding self-renewal in committed erythroid progenitors improves the biological relevance of a mathematical model of erythropoiesis. *J. Theor. Biol.* **250**, 322–338 (2008).
12. Chulián, S. *et al.* Dynamical properties of feedback signalling in B lymphopoiesis: A mathematical modelling approach. *J. Theor. Biol.* **522**, 110685 (2021).
13. Colijn, C. & Mackey, M. A mathematical model of hematopoiesis-I. Periodic chronic myelogenous leukemia. *J. Theor. Biol.* **237**, 117–132 (2005).
14. Kimmel, M. Stochasticity and determinism in models of hematopoiesis. *Adv. Exp. Med. Biol.* **844**, 119–152 (2014).
15. Catlin, S. N., Abkowitz, J. L. & Guttorp, P. Statistical Inference in a Two-Compartment Model for Hematopoiesis. *Biometrics* **57**, 546–553 (2001).
16. Stumpf, P. S., Arai, F. & MacArthur, B. D. Modeling Stem Cell Fates using Non-Markov Processes. *Cell Stem Cell* **28**, 187–190 (2021).
17. Xu, J. *et al.* Statistical inference for partially observed branching processes with application to cell lineage tracking of in vivo hematopoiesis. *Ann. Appl. Stat.* **13**, 2091–2119 (2019).
18. Etheridge, A. *Some Mathematical Models from Population Genetics*. vol. 2012 (Springer Berlin Heidelberg, 2011).
19. Bessonov, N., Pujo-Menjouet, L. & Volpert, V. Cell Modelling of Hematopoiesis. *Math. Model. Nat. Phenom.* **1**, 81–103 (2006).
20. Krinner, A., Roeder, I., Loeffler, M. & Scholz, M. Merging concepts - coupling an agent-based model of hematopoietic stem cells with an ODE model of granulopoiesis. *BMC Syst. Biol.* **7**, 117 (2013).
21. Laurenti, E. & Göttgens, B. From haematopoietic stem cells to complex differentiation landscapes. *Nature* **553**, 418–426 (2018).
22. Baey, C., Cournède, P.-H. & Kuhn, E. Asymptotic distribution of likelihood ratio test statistics for variance components in nonlinear mixed effects models. *Comput. Stat. Data Anal.* **135**, 107–122 (2019).
23. Hoekstra, R. & Vazire, S. Aspiring to greater intellectual humility in science. *Nat. Hum. Behav.* **5**, 1602–1607 (2021).
24. Duchesne, R., Guillemin, A., Crauste, F. & Gandrillon, O. Calibration, Selection and Identifiability Analysis of a Mathematical Model of the in vitro Erythropoiesis in Normal and Perturbed Contexts. *In Silico Biol.* **13**, 55–69 (2019).
25. Nocedal, J., & Wright, S. J. *Numerical Optimization*. (Springer-Verlag, 1999). doi:10.1007/b98874.
26. Hansen, N. *The CMA evolution strategy: a comparing review. Towards a new evolutionary computation*. (2006).
27. Andrieu, C. & Thoms, J. A tutorial on adaptive MCMC. *Stat. Comput.* **18**, 343–373 (2008).
28. Sisson, S. A., Fan, Y., & Beaumont, M. *Handbook of Approximate Bayesian Computation*. (Chapman and Hall/CRC, 2018). doi:10.1201/9781315117195.
29. Traulsen, A., Pacheco, J. M. & Dingli, D. On the Origin of Multiple Mutant Clones in Paroxysmal Nocturnal Hemoglobinuria. *Stem Cells* **25**, 3081–3084 (2007).

30. Haeno, H., Levine, R. L., Gilliland, D. G. & Michor, F. A progenitor cell origin of myeloid malignancies. *Proc. Natl. Acad. Sci.* **106**, 16616–16621 (2009).
31. Lyne, A.-M., Laplane, L. & Perié, L. To portray clonal evolution in blood cancer, count your stem cells. *Blood* **137**, 1862–1870 (2021).
32. Dingli, D. & Michor, F. Successful Therapy Must Eradicate Cancer Stem Cells. *Stem Cells* **24**, 2603–2610 (2006).
33. Lee-Six, H. *et al.* The landscape of somatic mutation in normal colorectal epithelial cells. *Nature* **574**, 532–537 (2019).
34. Welch, J. S. *et al.* The Origin and Evolution of Mutations in Acute Myeloid Leukemia. *Cell* **150**, 264–278 (2012).
35. Van Egeren, D. *et al.* Reconstructing the Lineage Histories and Differentiation Trajectories of Individual Cancer Cells in Myeloproliferative Neoplasms. *Cell Stem Cell* **28**, 514-523.e9 (2021).
36. Williams, N. *et al.* Life histories of myeloproliferative neoplasms inferred from phylogenies. *Nature* **602**, 162–168 (2022).
37. Watson, C. J. *et al.* The evolutionary dynamics and fitness landscape of clonal hematopoiesis. *Science* **367**, 1449–1454 (2020).
38. Stiehl, T. & Marciniak-Czochra, A. Mathematical Modeling of Leukemogenesis and Cancer Stem Cell Dynamics. *Math. Model. Nat. Phenom.* **7**, 166–202 (2012).
39. Moore, H. & Li, N. K. A mathematical model for chronic myelogenous leukemia (CML) and T cell interaction. *J. Theor. Biol.* **227**, 513–523 (2004).
40. Andersen, M. *et al.* Mathematical modelling as a proof of concept for MPNs as a human inflammation model for cancer development. *PLOS ONE* **12**, e0183620 (2017).
41. Bonnet, C. *et al.* Multistage hematopoietic stem cell regulation in the mouse: A combined biological and mathematical approach. *iScience* **24**, 103399 (2021).
42. Clapp, G. & Levy, D. A review of mathematical models for leukemia and lymphoma. *Drug Discov. Today Dis. Models* **16**, 1–6 (2015).
43. Michor, F. *et al.* Dynamics of chronic myeloid leukaemia. *Nature* **435**, 1267–70 (2005).
44. Roeder, I. *et al.* Dynamic modeling of imatinib-treated chronic myeloid leukemia: functional insights and clinical implications. *Nat. Med.* **12**, 1181–1184 (2006).
45. Foo, J., Drummond, M. W., Clarkson, B., Holyoake, T. & Michor, F. Eradication of Chronic Myeloid Leukemia Stem Cells: A Novel Mathematical Model Predicts No Therapeutic Benefit of Adding G-CSF to Imatinib. *PLoS Comput. Biol.* **5**, e1000503 (2009).
46. Glauche, I. *et al.* Therapy of chronic myeloid leukaemia can benefit from the activation of stem cells: simulation studies of different treatment combinations. *Br. J. Cancer* **106**, 1742–1752 (2012).
47. Bunimovich-Mendrazitsky, S., Kronik, N. & Vainstein, V. Optimization of Interferon–Alpha and Imatinib Combination Therapy for Chronic Myeloid Leukemia: A Modeling Approach. *Adv. Theory Simul.* **2**, 1800081 (2019).
48. Pedersen, R. K. *et al.* Dose-dependent mathematical modeling of interferon-α-treatment for personalized treatment of myeloproliferative neoplasms. *Comput. Syst. Oncol.* **1**, (2021).
49. Mosca, M. *et al.* Inferring the dynamics of mutated hematopoietic stem and progenitor cells induced by IFNα in myeloproliferative neoplasms. *Blood* **138**, 2231–2243 (2021).
50. Hermange, G., Vainchenker, W., Plo, I. & Cournède, P.-H. Mathematical modelling,

selection and hierarchical inference to determine the minimal dose in IFN alpha therapy against Myeloproliferative Neoplasms. *ArXiv211210688 Q-Bio Stat* (2021).
51. Hermange G., Rakotonirainy A., Bentriou M., et al. Inferring the initiation and development of myeloproliferative neoplasms. Proc Nat Ac Sci 2022;119(37);e2120374119.

Légendes des figures

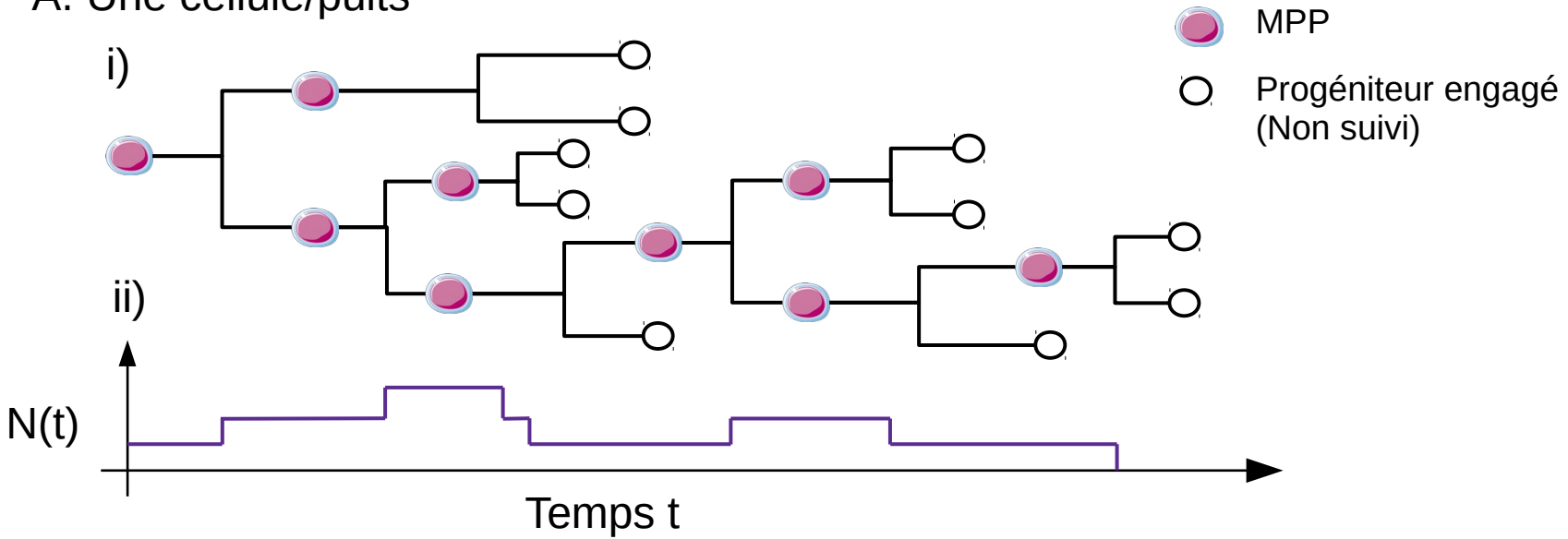
Figure 1 : Comparaison entre un modèle stochastique et son approximation déterministe
L'évolution du nombre de cellules hématopoïétiques d'une population donnée (considérons ici des progéniteurs multipotents – MPP) peut se modéliser par un processus de branchement à temps continu dans lequel un MPP se diviserait à un taux α = 0.5 jour^{-1}, et lorsqu'il se diviserait pourrait produire 2 MPP avec une probabilité p_2=0.2, 1 MPP et un progéniteur engagé (pour lequel on ne fait pas le suivi dans le temps des divisions) avec une probabilité p_1=0.3 ou 0 MPP avec une probabilité p_0=0.5.
A. Représentation d'une trajectoire aléatoire pour le processus stochastique considéré, partant d'un MPP mis en culture dans un puits à t=0. i) Schéma des branchements (divisions) et ii) évolution correspondante du nombre N(t) de MPP dans le puits au cours du temps.
B. Représentation de 10 trajectoires aléatoires simulées suivant le modèle. Ces 10 trajectoires peuvent être interprétées comme l'évolution du nombre de MPP dans 10 puits dans lesquels on aurait mis 50 MPP initialement. Le modèle étant stochastique, pour des valeurs données des paramètres, on observe différentes dynamiques.
C. Lorsque le nombre de cellules devient très grand, la variabilité devient négligeable et on peut faire une approximation déterministe du modèle. Pour des valeurs données des paramètres, une seule dynamique est alors produite par le modèle, correspondant à la trajectoire de la population.

Figure 2 : Illustration de la méthodologie sur un exemple fictif
A. Considérons une expérience fictive dans laquelle on mesure une concentration en MPP y(t) à différents instants t (en jours). Les données sont bruitées. L'échantillon à 4 jours est l'échantillon contrôle. On souhaite prédire la dynamique de diminution du nombre de MPP au delà de 3 jours.
B. Pour répondre à cet objectif, on construit un modèle mathématique à un compartiment, modélisé par une équation différentielle ordinaire. La solution de cette équation est une loi exponentiel (modèle exponentiel). Il n'est pas possible d'estimer à la fois α, p_2 et p_0 (non-identifiabilité du modèle). En définissant $K=\alpha(p_0-p_2)$, le modèle devient identifiable.
C. Pour estimer la valeur de $K_{réel}$, on minimise une distance (l'erreur quadratique moyenne – mse) entre les observations et les valeurs du modèle, en fonction de la valeur de K. On trouve comme valeur optimale $K_{estimé}$ = 0.59 et comme intervalle de confiance à 95% : $K_{estimé} \in$ [0.56,0.62].
D. On peut comparer les résultats du modèle exponentiel (mse : e=5.0e-4) avec ceux qu'on obtiendrait par une régression linéaire (e=6.0e-3) ou polynomiale (e=4.3e-4). On obtient des résultats légèrement meilleurs avec un modèle polynomial plutôt qu'avec le modèle exponentiel, mais le polynôme fait aussi intervenir un paramètre en plus : il est moins parcimonieux. Pour valider le modèle exponentiel, on peut alors comparer l'erreur de prédiction sur l'observation à 4 jours et constater qu'il permet en effet de bien prédire la concentration en MPP, contrairement aux modèles polynomial et linéaire.
Notons que sur cet exemple très simple, les données avaient été générées par le modèle exponentiel avec $K_{réel}$=0.6 et un bruit additif Gaussien d'écart-type σ=0.05.

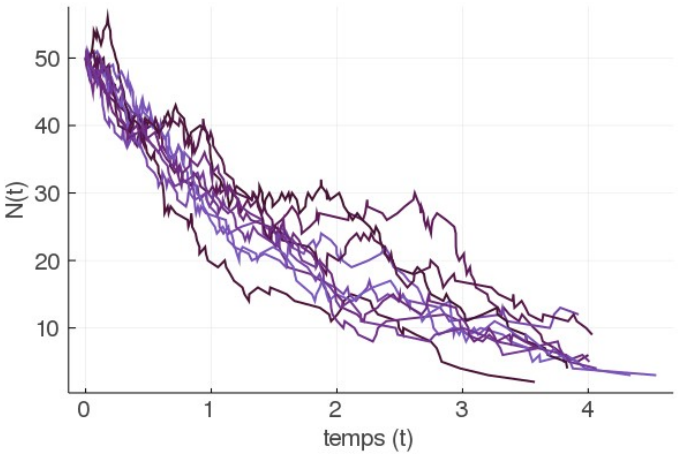
Figure 3 : Intérêt des modèles mathématiques
Différents intérêts (liste non exhaustive) des modèles mathématiques.
A. Les modèles mathématiques peuvent permettre d'inférer des processus latents non directement observables, tels que la dynamique au niveau de cellules souches (CSH) ou l'histoire de l'apparition et du développement des hémopathies malignes, ce qui peut permettre par la suite la mise en place de méthodes de dépistage précoce.
B. Les modèles mathématiques peuvent inférer des mécanismes biologiques, en testant par exemple différentes hypothèses biologiques, ce qui peut servir à identifier des pistes thérapeutiques et à construire des essais cliniques qui permettront leur validation et leur usage en clinique.
C. Les modèles mathématiques sont des outils d'aide à la décision clinique, permettant de comprendre le mécanisme d'action des traitements, de prédire leurs effets et d'optimiser leur posologie.

Figure 1

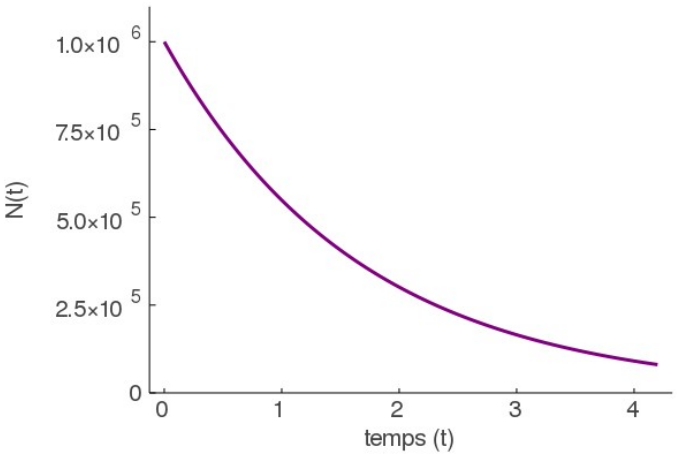
A. Une cellule/puits



B. 50 cellules/puits



C. Un million de cellules/puits



<u>Figure 1</u> : Comparaison entre un modèle stochastique et son approximation déterministe

L'évolution du nombre de cellules hématopoïétiques d'une population donnée (considérons ici des progéniteurs multipotents – MPP) peut se modéliser par un processus de branchement à temps continu dans lequel un MPP se diviserait à un taux $\alpha = 0.5$ jour^{-1}, et lorsqu'il se diviserait pourrait produire 2 MPP avec une probabilité $p_2=0.2$, 1 MPP et un progéniteur engagé (pour lequel on ne fait pas le suivi dans le temps des divisions) avec une probabilité $p_1=0.3$ ou 0 MPP avec une probabilité $p_0=0.5$.

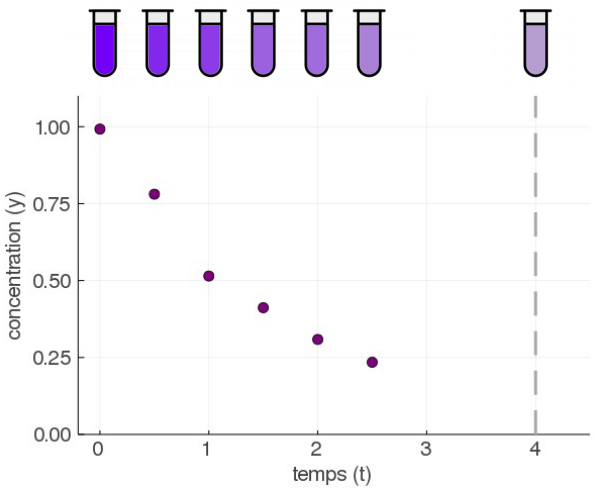
A. Représentation d'une trajectoire aléatoire pour le processus stochastique considéré, partant d'un MPP mis en culture dans un puits à t=0. i) Schéma des branchements (divisions) et ii) évolution correspondante du nombre N(t) de MPP dans le puits au cours du temps.

B. Représentation de 10 trajectoires aléatoires simulées suivant le modèle. Ces 10 trajectoires peuvent être interprétées comme l'évolution du nombre de MPP dans 10 puits dans lesquels on aurait mis 50 MPP initialement. Le modèle étant stochastique, pour des valeurs données des paramètres, on observe différentes dynamiques.

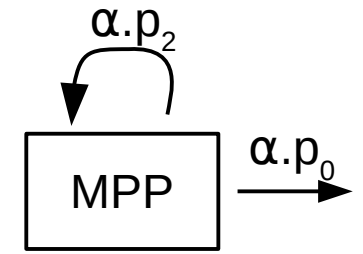
C. Lorsque le nombre de cellules devient très grand, la variabilité devient négligeable et on peut faire une approximation déterministe du modèle. Pour des valeurs données des paramètres, une seule dynamique est alors produite par le modèle, correspondant à la trajectoire de la population.

Figure 2

A. Données

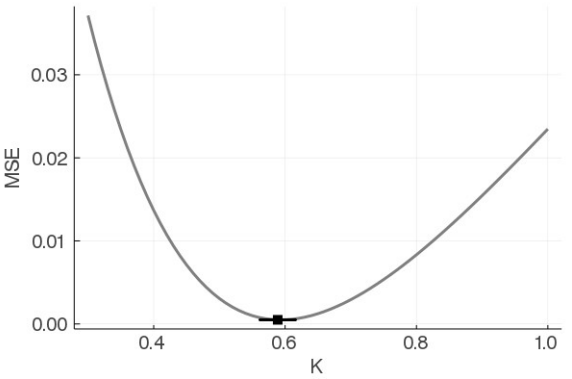


B. Modèle

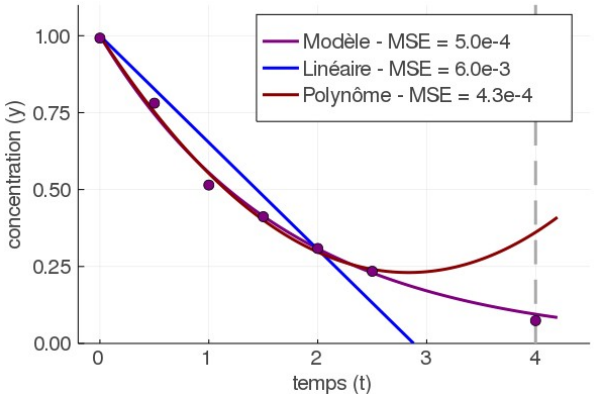


$$\dot{y}(t) = \alpha(p_2 - p_0)y(t)$$
$$= -Ky(t)$$

C. Estimation



D. Validation



<u>Figure 2</u> : Illustration de la méthodologie sur un exemple fictif

A. Considérons une expérience fictive dans laquelle on mesure une concentration en MPP *y(t)* à différents instants *t* (en jours). Les données sont bruitées. L'échantillon à 4 jours est l'échantillon contrôle. On souhaite prédire la dynamique de diminution du nombre de MPP au delà de 3 jours.

B. Pour répondre à cet objectif, on construit un modèle mathématique à un compartiment, modélisé par une équation différentielle ordinaire. La solution de cette équation est une loi exponentiel (modèle exponentiel). Il n'est pas possible d'estimer à la fois α, p_2 et p_0 (non-identifiabilité du modèle). En définissant $K=\alpha(p_0-p_2)$, le modèle devient identifiable.

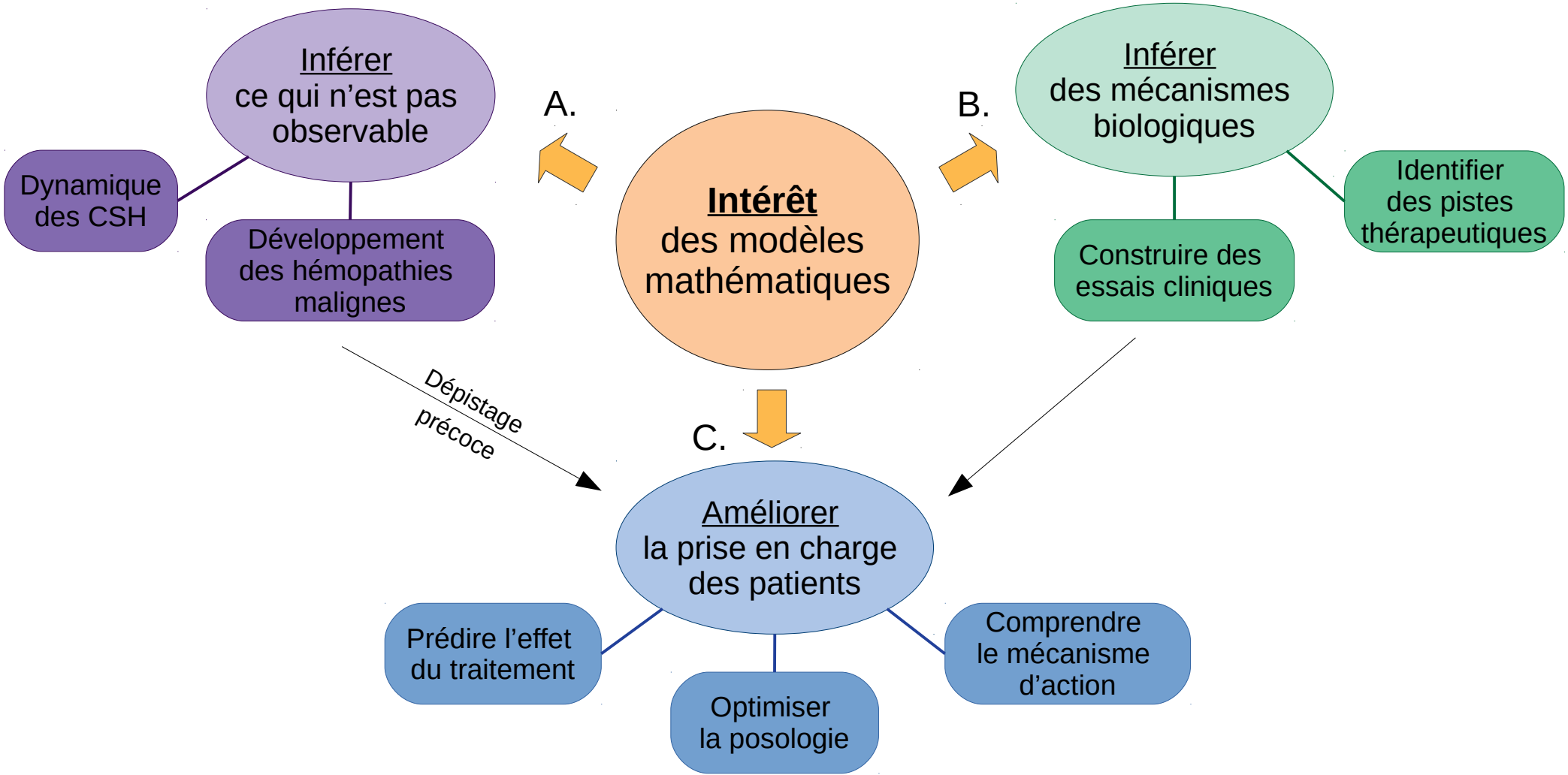
C. Pour estimer la valeur de $K_{réel}$, on minimise une distance (l'erreur quadratique moyenne – mse) entre les observations et les valeurs du modèle, en fonction de la valeur de K. On trouve comme valeur optimale $K_{estimé} = 0.59$ et comme intervalle de confiance à 95% : $K_{estimé} \in [0.56,0.62]$.

D. On peut comparer les résultats du modèle exponentiel (mse : $e=5.0e^{-4}$) avec ceux qu'on obtiendrait par une régression linéaire ($e=6.0e^{-3}$) ou polynomiale ($e=4.3e^{-4}$). On obtient des résultats légèrement meilleurs avec un modèle polynomial plutôt qu'avec le modèle exponentiel, mais le polynôme fait aussi intervenir un paramètre en plus : il est moins parcimonieux.

Pour valider le modèle exponentiel, on peut alors comparer l'erreur de prédiction sur l'observation à 4 jours et constater qu'il permet en effet de bien prédire la concentration en MPP, contrairement aux modèles polynomial et linéaire.

Notons que sur cet exemple très simple, les données avait été générées par le modèle exponentiel avec $K_{réel}=0.6$ et un bruit additif Gaussien d'écart-type $\sigma=0.05$.

Figure 3


Inférer
ce qui n'est pas
observable
Dynamique
des CSH
Développement
des hémopathies
malignes
A.
Intérêt
des modèles
mathématiques
B.
Inférer
des mécanismes
biologiques
Identifier
des pistes
thérapeutiques
Construire des
essais cliniques
Dépistage
précoce
C.
Améliorer
la prise en charge
des patients
Prédire l'effet
du traitement
Optimiser
la posologie
Comprendre
le mécanisme
d'action

<u>Figure 3</u> : Intérêt des modèles mathématiques

Différents intérêts (liste non exhaustive) des modèles mathématiques.

A. Les modèles mathématiques peuvent permettre d'inférer des processus latents non directement observables, tels que la dynamique au niveau de cellules souches (CSH) ou l'histoire de l'apparition et du développement des hémopathies malignes, ce qui peut permettre par la suite la mise en place de méthodes de dépistage précoce.

B. Les modèles mathématiques peuvent inférer des mécanismes biologiques, en testant par exemple différentes hypothèses biologiques, ce qui peut servir à identifier des pistes thérapeutiques et à construire des essais cliniques qui permettront leur validation et leur usage en clinique.

C. Les modèles mathématiques sont des outils d'aide à la décision clinique, permettant de comprendre le mécanisme d'action des traitements, de prédire leurs effets et d'optimiser leur posologie.