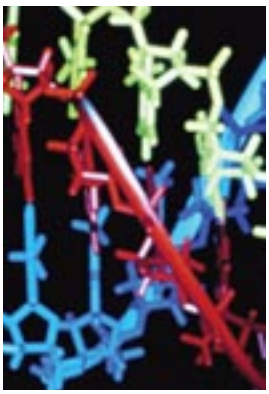
MEDECINE/SCIENCES 2002 ; 18 : 492-502

Bio-informatique (3)

Modélisation, analyse et simulation des réseaux génétiques

Denis Thieffry, Hidde De Jong

> A l'interface de la biologie, de l'informatique et des mathématiques, les méthodes de modélisation, d'analyse et de simulation permettent de décrire et de prédire le comportement de réseaux de régulation génétique complexes. Ces méthodes se déclinent en quatre principales catégories en fonction des notions mathématiques utilisées : le recours aux concepts de la théorie des graphes, l'utilisation d'équations différentielles, les formalisations booléennes ou logique généralisée, ou encore une description stochastique des transitions entre états moléculaires. Au vu des données génétiques et moléculaires actuellement disponibles, les approches qualitatives (éventuellement sur base de modèles quantitatifs) semblent particulièrement indiquées. Encore en cours de développement, de telles approches et leur automatisation devraient rapidement devenir incontournables pour caractériser les comportements dynamiques normaux ou pathologiques des réseaux biologiques, pour prédire les effets de perturbations, ou encore pour concevoir une nouvelle génération d'outils thérapeutiques ou agronomiques. <



D. Thieffry : Laboratoire de Génétique et Physiologie du Développement, Parc Scientifique de Luminy, 13288 Marseille Cedex 9, France.
thieffry@lgpd.univ-mrs.fr
H. De Jong : Inria Rhône-Alpes, 655, avenue de l'Europe, Montbonnot, 38334 Saint Ismier Cedex, France.
Hidde.de-Jong@inrialpes.fr

Les récents progrès technologiques en biologie ont conduit à un véritable changement d'échelle en ce qui concerne la caractérisation de la composition moléculaire des êtres vivants. En effet, les techniques de séquençage, à présent largement automatisées, ont déjà mené au déchiffrement de plusieurs dizaines de génomes complets, en majorité des bactéries, mais aussi plusieurs génomes d'organismes pluricellulaires comme le nématode *C. elegans*, la mouche *D. melanogaster*, ou encore la plante modèle *A. thaliana*. Le génome humain est, quant à lui, déjà en cours de polissage et d'assemblage. Les génomes de souris et de divers autres organismes modèles ne devraient plus tarder à suivre.

Par ailleurs, de nouvelles techniques de caractérisation à grande échelle des états d'expression génique (méthode SAGE, puces à ADN, gels de protéines 2D couplés à la spectrométrie de masse, ...), ou encore des capacités interactives des protéines (puces à ADN double-brin, cribles simple, double, ou triple hybride...), comblent maintenant partiellement le fossé entre la production massive de données structurales sur les génomes (séquences) et la caractérisation fonctionnelle des nombreuses macromolécules ainsi identifiées [1-7].

Nous avons à notre disposition des catalogues de gènes assez complets pour un certain nombre d'organismes modèles. Le recoupement de données expérimentales diverses et l'exploitation des outils d'analyse et de comparaison de séquences permettent souvent de définir, au moins partiellement, les propriétés fonctionnelles des protéines codées par ces gènes. Cependant, dans la plupart des cas, il reste très difficile d'évaluer précisément les mécanismes régulateurs qui président à l'expression différentielle des gènes, même dans le cas d'organismes modèles étudiés depuis des décennies. Dans la mesure où ces mécanismes de régulation sont

en grande partie à l'origine des processus de différenciation au cours du développement embryonnaire, de la diversité cellulaire et tissulaire, ainsi que des capacités d'adaptation et de réaction physiologique des organismes supérieurs, il est crucial de surmonter de telles limitations dans notre capacité de caractérisation fonctionnelle des génomes.

Vers la caractérisation fonctionnelle des réseaux moléculaires à l'échelle des génomes

Au niveau des séquences génomiques elles-mêmes, de sérieux progrès restent à accomplir quant à la caractérisation précise des régions transcrites et, surtout, des régions *cis*-régulatrices, parfois plus grandes et plus complexes que les régions codantes dont elles contrôlent la transcription.

Des revues récentes fondées sur quelques dizaines de gènes étudiés en détail suggèrent qu'un gène du développement peut être contrôlé par des dizaines de facteurs de régulation, chacun susceptible de se fixer au niveau de plusieurs sites, formant des régions fonctionnant de manière intégrée (régions promotrices et *enhancer*) [8]. Différentes combinaisons de facteurs peuvent ainsi conduire à des états d'expression différents, en fonction de la région et du stade embryonnaire. La prédiction individuelle des promoteurs ou des sites de régulation transcriptionnelle est rendue difficile par la petite taille et/ou la faible conservation des régions d'ADN impliquées. La plupart des chercheurs s'accordent sur la nécessité de caractériser des combinaisons fonctionnelles de sites de fixation pour les protéines, mais aucun outil n'est encore réellement opérationnel pour cette tâche.

Outre la caractérisation des régions *cis*-régulatrices, et donc des divers facteurs impliqués dans le contrôle de l'expression d'un gène donné, il s'agit également de comprendre les propriétés dynamiques collectives des groupes de gènes qui se contrôlent les uns les autres. De telles régulations croisées peuvent former des hiérarchies complexes conditionnant l'expression d'ensembles de gènes à celle d'autres, mais elles peuvent aussi former des réseaux imbriqués conduisant à de riches propriétés dynamiques, par exemple des états d'expression périodique, ou encore des états d'expression alternatifs multiples. La *Figure 1* présente un schéma synthétisant les interactions moléculaires (protéine-protéine, protéine-ADN, etc.) entre des facteurs impliqués dans des processus de transformation maligne chez les mammifères [9]. Bien que ce schéma ne constitue encore qu'une description partielle et simplifiée du réseau contrôlant la division, la mort programmée et la différenciation cellulaire, il est clair que notre intuition est désarmée face à un tel écheveau interactif (voir aussi [10]). Pourtant, une compréhension fine du comportement global d'un tel réseau est indispensable pour pouvoir diagnostiquer, prédire, voire corriger les dérèglements conduisant au cancer.

C'est précisément à la description de méthodes permettant l'intégration, l'analyse et la simulation de tels réseaux de régulation qu'est consacré cet article. Malheureusement, à l'interface entre la biologie, l'informatique et les mathématiques, ces méthodes font appel à des outils théoriques qui apparaîtront parfois

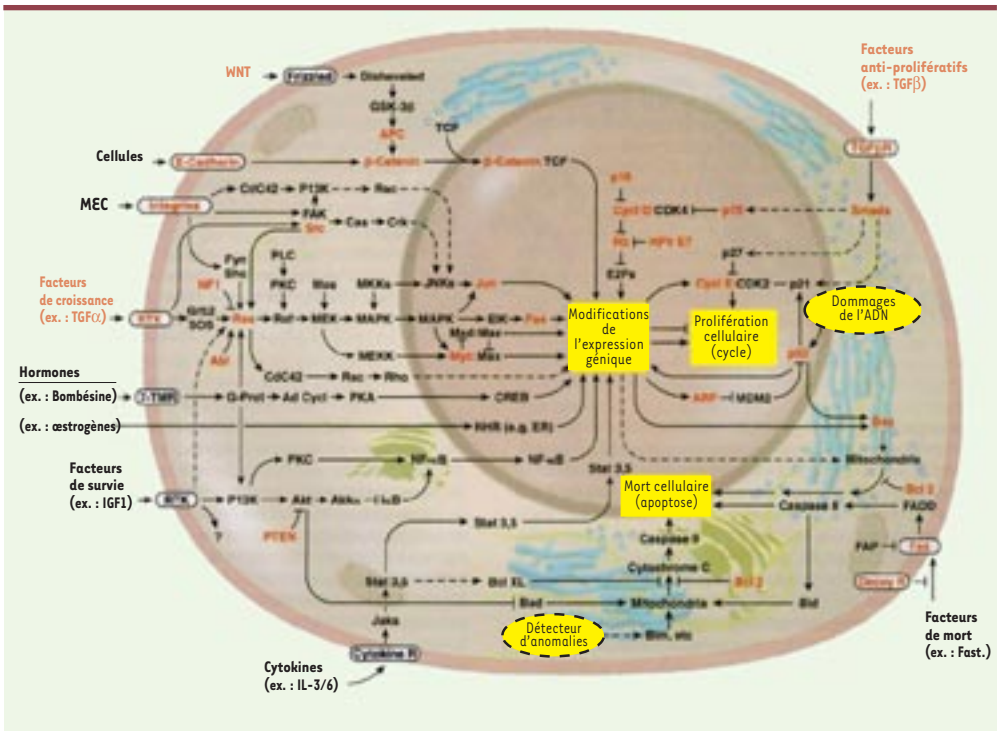


Figure 1. ***Représentation schématique du réseau de régulation formé par les principaux facteurs impliqués dans le contrôle du cycle ou de la mort cellulaire chez les mammifères.*** La plupart de ces facteurs sont impliqués dans l'une ou l'autre étape de transformation cancéreuse (reproduit de *Cell* [9], avec la permission d'Elsevier Science).

ardus ou abstraits au biologiste. Nous éviterons donc d'entrer dans trop de détails formels pour nous concentrer sur les stratégies et les concepts qui sont à la base des différentes approches en cours de développement.

Le processus de modélisation

D'un point de vue général, le processus de modélisation dynamique d'un réseau de régulation requiert :

1. L'identification des éléments pertinents (gènes et leurs produits, signaux moléculaires, etc.). Une telle identification présuppose le choix d'un niveau de représentation : s'agit-il de décrire l'évolution de la concentration de toutes les espèces moléculaires en présence : ARNm, protéines, métabolites, etc ? Ou désire-t-on plutôt rendre compte seulement de l'existence de plusieurs états d'expression alternatifs, ainsi que des voies menant à ces états ? Dans ce dernier cas, il suffira d'introduire autant de variables que de gènes impliqués, les détails moléculaires étant donc traités de manière implicite.

2. L'identification des interactions entre les éléments considérés. À nouveau, cette identification dépendra du niveau de description. Dans certains cas, on pourra se contenter de décrire les influences régulatrices entre gènes tandis que, dans d'autres cas, il conviendra de détailler les mécanismes moléculaires impliqués dans la régulation des gènes.

3. Le choix de fonctions mathématiques pour représenter une interaction ou un ensemble d'interactions impliquées par exemple dans la régulation de l'expression d'un gène.

4. Le cas échéant, l'évaluation des valeurs ou de domaines de valeurs pour les paramètres présents dans les équations, soit en fonction des données expérimentales disponibles, soit au regard de considérations théoriques.

Ces quatre points ne constituent pas à proprement parler une série d'étapes ordonnées temporellement. Ils sont généralement abordés en parallèle, et bien souvent de manière itérative. En fait, au fil des simulations ou de la résolution des équations produites et de la confrontation des résultats obtenus aux données expérimentales, on est souvent conduit à réévaluer l'un ou l'autre de ces points, jusqu'à l'obtention d'un comportement dynamique cohérent avec les données disponibles. Ce processus de modélisation et ses composantes ne sont, bien entendu, pas propres à la biologie. On pourra ainsi les retrouver dans tout processus de modélisation dynamique, depuis les sciences « exactes » jusqu'aux sciences humaines (par exemple en économie).

Dans le cas de la biologie, on est cependant confronté à plusieurs limitations importantes. D'une part, à l'échelle d'un organisme, le nombre d'éléments différents pertinents peut être très grand. De plus, ces éléments peuvent interagir avec de nombreux autres éléments, et ce, de manière variée, parfois sensible au contexte (intervention de co-facteurs, influence de l'ordre de succession des événements...). Enfin, la nature de ces interactions et de leurs effets est le plus souvent connue de manière plutôt qualitative. Il s'avère donc en général difficile de justifier rigoureusement l'usage de l'une ou l'autre fonction mathématique quantitative, et plus encore de déterminer précisément les valeurs paramétriques pertinentes à partir des données expérimentales. Quels que soient les outils mathématiques utilisés, il sera crucial de pouvoir apprécier l'ensemble des propriétés dynamiques des modèles obtenus, ainsi que les conditions paramétriques ou structurelles (fonctions utilisées) associées.

D'un point de vue formel, on peut répartir les approches mathématiques pour la modélisation des réseaux molé-

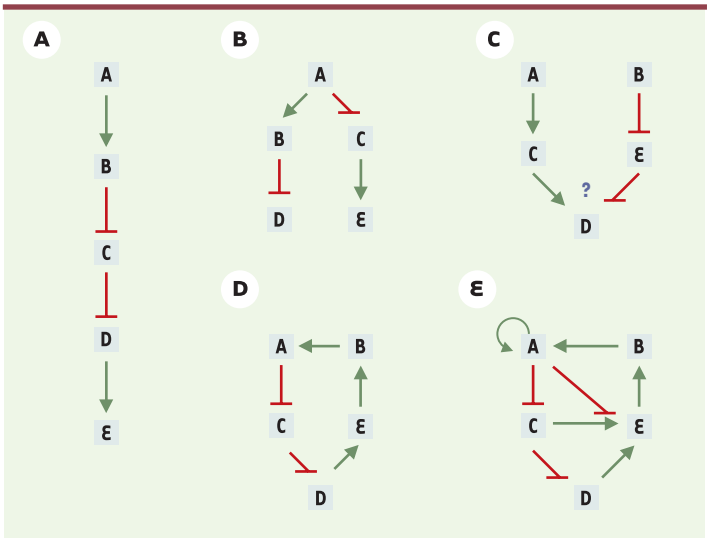


Figure 2. ***Quelques exemples de graphes d'interaction.*** Au sein de ces graphes, les sommets font référence aux gènes. Les interactions entre les gènes A-E sont représentées par des arcs terminés par une flèche verte (activations) ou par un trait perpendiculaire rouge (inhibitions). Les détails moléculaires à la base de ces interactions sont implicites (transcriptions, traductions, associations et interactions entre macromolécules, etc.). Un arc représente par conséquent l'effet global (positif ou négatif) de l'expression d'un gène sur celle d'un autre gène. Dans les cas ***A***, ***B*** et ***C***, il est facile de prédire les répercussions d'une surexpression ou d'une perte de fonction d'un gène à quelque niveau que ce soit dans la hiérarchie. Le cas ***D*** représente une cascade de régulation bouclée sur elle-même : l'expression de chaque gène, directement influencée par un seul gène et influençant directement un seul autre gène, dépend aussi indirectement de sa propre expression. Enfin, le cas ***E*** comporte plusieurs circuits imbriqués, ce qui complique sérieusement toute analyse intuitive.

culaires biologiques en quatre catégories principales, faisant respectivement appel à la théorie des graphes, à une formalisation discrète ou logique, aux systèmes d'équations différentielles, ou encore à des équations stochastiques (voir [11, 12] pour des revues récentes). Nous allons maintenant brièvement introduire ces différents types de formalisations pour ensuite discuter leurs limites et avantages respectifs.

Des schémas de régulation aux graphes des mathématiciens

Le premier type de formalisation est proche de la représentation schématique graphique à laquelle le biologiste a l'habitude de recourir, au moins dans le cadre de présentations synthétiques des informations obtenues sur les éléments et les interactions du système biologique étudié. D'un point de vue formel, un graphe est défini par les « sommets » qui le composent, auxquels s'ajoutent des « arêtes » reliant des paires de sommets. Ces arêtes peuvent être orientées (on parle alors d'« arcs ») et signées (signe moins dans le cas d'une inhibition, signe plus dans le cas d'une activation). D'un point de vue biologique, les sommets peuvent par exemple représenter des gènes de régulation, et les arêtes des interactions entre ces gènes *via* leurs produits (ARNm ou protéine). Afin de pouvoir exploiter les concepts, les outils d'analyse et les résultats mathématiques de la théorie des graphes, il s'agit de définir rigoureusement les règles de représentation des éléments et des interactions biologiques en un langage simplifié, standardisé, et la plupart du temps univoque (voir [13-15] pour une application de ces concepts à l'analyse des réseaux métaboliques). Quelques exemples de graphes représentant des réseaux d'interaction sont repris dans la *Figure 2*. Les règles de représentation choisies attribuent un sommet par gène et distinguent entre des interactions régulatrices positives ou négatives (arcs signés).

Pour des cas simples, une telle représentation sous forme de graphe permet déjà de répondre à diverses questions biologiques, par exemple prédire les répercussions d'une surexpression ou de la perte de fonction d'un gène à quelque niveau que ce soit dans la hiérarchie *(Figure 2A* et *B)*. En revanche, quand deux séquences d'interactions convergent vers un même gène *(Figure 2C)*, il devient nécessaire de préciser la manière dont la co-régulation s'exerce sur ce gène pour prévoir les conséquences d'une mutation en aval. L'absence d'inhibiteur suffit-elle pour que ce gène soit exprimé ? Faut-il juste la présence de l'activateur, même en présence du répresseur ? Ou faut-il à la fois l'absence du répresseur et la présence de l'activateur ? Dès que les cascades, convergentes ou divergentes, commencent à s'entremêler en formant des réseaux avec rétroactions, il devient très difficile de prédire le comportement du système. C'est le cas notamment des cascades de régulation bouclées sur elles-mêmes *(Figure 2D)*. De tels « boucles » ou « circuits » de régulation constituent de véritables cercles vicieux, où l'expression de chaque gène, directement influencée par un seul gène et influençant directement un seul autre gène, dépend aussi indirectement de sa propre expression préalable *via* les autres gènes du circuit. D'un point de vue théorique, il est possible de répartir ces circuits de régulation en deux grandes classes dotées de propriétés bien définies [16]. Au sein de la première catégorie de circuits, chaque élément a un effet indirect positif sur lui-même et l'on parlera alors de « circuit positif ». De tels circuits peuvent produire deux états d'expression alternatifs (« attracteurs »), en général diamétralement différents. On parle alors de « multistationnarité » d'un point de vue mathématique, ou de « différenciation » en biologie. De tels circuits

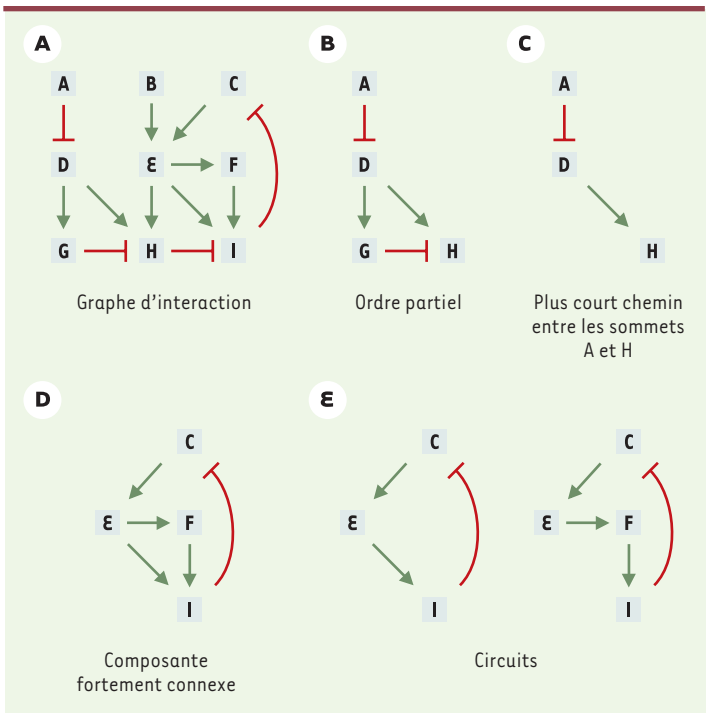


Figure 3. ***Illustration de quelques notions de la théorie des graphes.***
A. Exemple de graphe d'interactions complexe. Il est possible d'extraire de ce graphe plusieurs sous-graphes. ***B.*** Sous-graphe présentant un ordre partiel. ***C.*** Sous-graphe présentant le plus court chemin entre deux sommets. ***D.*** Composante fortement connexe : partant d'un élément, il existe au moins un cycle qui ramène vers cet élément. ***E.*** Cycle orienté : il s'agit d'un type particulier de composante fortement connexe qui ne possède qu'un seul cycle. Les arcs se terminant par une flèche représentent des activations, les arcs terminés par un trait perpendiculaire représentent des inhibitions impliquant les gènes A à I.

permettent de mémoriser de manière durable un signal intercellulaire ou environnemental sous la forme d'un état d'expression différencié [17, 18]. Dans le cas de la seconde catégorie de circuits, chaque élément a un effet indirect négatif sur lui-même et l'on parle alors de « circuit négatif ». De tels circuits peuvent donner lieu à une expression homéostatique, éventuellement périodique, pour tous les gènes impliqués. Nous verrons plus loin des exemples simples pour les deux catégories de circuits dans le cadre d'une formalisation dynamique plus explicite.

Dès que plusieurs circuits sont imbriqués *(Figure 2E)*, il devient très difficile d'évaluer les propriétés dynamiques du réseau. En fait, en fonction des contraintes existantes sur l'expression des différents gènes impliqués, plusieurs cas de figure peuvent se présenter, depuis la prédominance d'un seul circuit jusqu'à la combinaison des propriétés de plusieurs circuits. Il faut alors passer à une formalisation dynamique explicite (équations différentielles, logiques ou stochastiques, *voir plus loin*) pour préciser les différents comportements dynamiques possibles et les conditions paramétriques associées.

Avant de nous tourner vers ces formalisations dynamiques, il est important de réaliser que la théorie des graphes offre déjà un ensemble de concepts et d'algorithmes permettant d'aborder de nombreuses questions biologiques. En particulier, il est possible de décomposer un graphe complexe en composantes plus facilement interprétables *(Figure 3)*. La théorie des graphes devrait aussi permettre d'envisager de manière formelle et générique les problèmes de comparaison entre sous-réseaux, au sein d'un organisme, ou en comparant les interactions entre gènes dans des organismes différents. On parle alors d'« isomorphisme » (même structure : même nombre d'éléments reliés par des configurations d'arêtes équivalentes) ou encore d'« homéomorphisme » (même topologie, par exemple en termes de circuits) entre graphes.

Modèles différentiels

Parmi les méthodes de modélisation dynamique, la plus utilisée en biologie est sans conteste la description différentielle (voir par exemple [19-25]). Les concentrations ou les activités des espèces moléculaires sont généralement représentées par des grandeurs (ou « variables ») réelles positives, susceptibles de varier

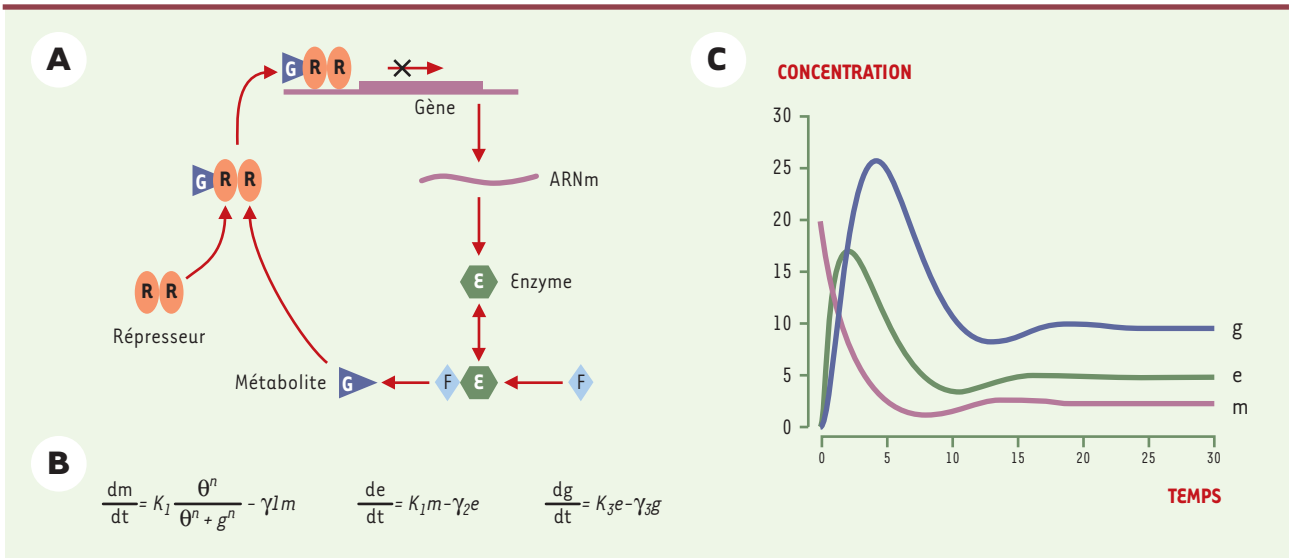


Figure 4. ***Modélisation différentielle du réseau correspondant au cas classique d'inhibition d'une réaction par son produit*** (d'après [26]). ***A.*** Schéma réactionnel. Ε (pour « enzyme ») et R (pour répresseur) représentent des protéines, et F et G représentent des métabolites. L'association du répresseur (sous forme de dimère) au métabolite G permet la fixation du complexe en amont du gène codant pour l'enzyme Ε et le blocage de la transcription de celui-ci. ***B.*** Système d'équations différentielles correspondant : m, e et g représentent les concentrations de l'ARNm codant pour l'enzyme, de la protéine enzymatique, et du métabolite G (co-répresseur), respectivement. k_m, k_e et k_g sont des constantes de synthèses, alors que γ_m, γ_e et γ_g sont des constantes de dégradation, θ est une constante de seuil, et n une constante de coopérativité. ***C.*** Résultat d'une simulation pour un choix de conditions initiales et des valeurs paramétriques raisonnables mais arbitraires. Les trois courbes correspondent aux nombres de molécules d'ARNm (variable m, violet), de protéine enzymatique (variable e, vert), et du métabolite G (co-répresseur, variable g, bleu). On obtient un état stationnaire stable.

de manière continue au cours du temps. La variation de ces grandeurs est formalisée par l'écriture d'un système d'équations différentielles couplées. Si l'on ne tient compte que des espèces moléculaires et de leurs interactions, on a un système d'équations différentielles ordinaires. Si l'on veut également prendre en compte explicitement la dimension spatiale et la diffusion des molécules impliquées, on se tourne alors vers l'utilisation d'équations aux dérivées partielles. Dans la plupart des situations biologiques, les interactions considérées sont non linéaires (effets de seuil, de saturation, interactions synergiques, etc.), ce qui conduit à des modèles différentiels généralement impossibles à résoudre de façon analytique. Il est alors très difficile de dériver les solutions de ces systèmes et l'on doit recourir à des « simulations numériques ». Partant de conditions initiales, il s'agit d'approcher la solution exacte par un calcul des valeurs des concentrations des espèces moléculaires impliquées au cours du temps, en utilisant des intervalles de temps arbitrairement petits. Il faut alors fixer les valeurs de tous les paramètres. Celles-ci n'étant pas toujours établies expérimentalement, les simulations sont effectuées avec des valeurs très approximatives, voire arbitraires. Il est par conséquent en général difficile d'établir le caractère représentatif du comportement dynamique prédit. Des techniques d'analyse numérique plus sophistiquées permettent de vérifier ce qu'il advient des états stationnaires lorsque l'on modifie la valeur de l'un ou l'autre paramètre. Néanmoins, pour des modèles biologiques un peu complexes, il est ardu de se faire une idée précise de l'ensemble des propriétés dynamiques correspondantes, ainsi que des conditions paramétriques associées.

Un exemple de modèle très simple est donné dans la *Figure 4*. Inspiré du travail pionnier de Brian Goodwin, ce modèle décrit un processus de régulation, très fréquent dans le cadre du métabolisme, où le produit d'une réaction participe à l'inhibition de l'expression d'un gène codant pour une enzyme qui catalyse une étape de cette même réaction [26]. Il s'agit donc d'un circuit de régulation négatif qui peut être représenté sous la forme d'un système de trois équations différentielles ordinaires et donner lieu à une simulation numérique *(Figure 4C)*. Pour les valeurs de paramètres choisies, le système évolue par des oscillations amorties vers un seul état stationnaire, stable. Comme nous l'avons suggéré plus haut, un tel comportement périodique (entretenu ou amorti) est caractéristique des circuits de régulation négatifs.

De nombreux auteurs ont proposé et simulé des modèles différentiels pour des réseaux moléculaires, impliqués par exemple dans la régulation de l'expression de gènes bactériens ou viraux [19], dans le contrôle du cycle cellulaire [20], dans la formation des rythmes circadiens [21] ou encore de profils d'expression géniques spécifiques au cours du développement embryonnaire [22-24], et ce, chez divers organismes modèles.

Modèles stochastiques

Le recours à des équations différentielles présuppose que les concentrations des molécules impliquées varient de manière continue et déterministe. Ces considérations peuvent s'avérer problématiques lorsque certaines molécules sont présentes en petit nombre. Ce peut être le cas pour certains facteurs de transcription, et en tous cas pour les gènes eux-mêmes qui sont généralement présents en un ou deux exemplaires dans les cellules. Par ailleurs, différents types de fluctuations peuvent affecter le déroulement temporel de nombreux processus moléculaires, par exemple le temps nécessaire à la transcription d'un gène. Pour prendre en compte de tels effets, plusieurs auteurs ont proposé des modèles dits stochastiques, formalisant l'évolution d'un système par des transitions entre différents états, chaque transition étant affectée d'une probabilité définie.

Une approche rigoureuse se fonde sur des équations appelées « équations maîtresses ». Malheureusement, il s'avère que la plupart du temps ces équations sont impossibles à résoudre analytiquement. Par conséquent, ici aussi, on fait appel à des méthodes de simulation qui permettent une approximation des solutions des équations maîtresses. Un tel exemple est présenté dans la *Figure 5*. En partant de l'ensemble des réactions qui peuvent se produire dans le cas du mécanisme d'inhibition par le substrat, l'évolution de l'état du système, c'est-à-dire du nombre de molécules de chaque type, est prédite. L'évolution est déterminée par des variables stochastiques représentant l'intervalle de temps entre deux réactions successives, ainsi que le type de la prochaine réaction. Avec des conditions initiales équivalentes à celles utilisées pour la simulation différentielle *(Figure 4C)*, l'allure des courbes de concentration pour les trois espèces moléculaires ressemble à celle des courbes différentielles, moyennant l'ajout d'un « bruit ». Le comportement prédit par le modèle différentiel peut en effet être interprété comme la moyenne des comportements prédits par le modèle stochastique. Ceci ne pose pas de problème dans ce cas précis où, étant donné le mécanisme de régulation du système, tous les comportements stochastiques suivent approximativement le comportement déterministe. Dans d'autres cas, toutefois, des divergences peuvent apparaître entre différentes simulations stochastiques

du même système, représentant par exemple des voies de développement alternatives pour la cellule. Les modèles différentiels sont incapables de prendre en compte explicitement de tels effets stochastiques.

Assez lourdes à mettre en œuvre, les simulations stochastiques n'ont été jusqu'à présent appliquées qu'à un petit nombre de réseaux moléculaires biologiques bien caractérisés expérimentalement, dont la régulation de l'expression du bactériophage lambda [27] et la cascade de phosphorylations impliquée dans un processus de chimiotaxie bactérienne [28].

Modèles qualitatifs

Malheureusement, il est généralement très difficile d'évaluer précisément les valeurs des principaux paramètres impliqués dans les modèles différentiels ou stochastiques de réseaux moléculaires. Potentiellement très précises, les expériences *in vitro* présentent l'inconvénient de ne pas correspondre aux conditions réelles au sein de la cellule ou de l'organisme étudié. Les mesures obtenues *in vivo* sont souvent d'une précision inférieure et peuvent provenir de plusieurs effets difficiles à distinguer. Même les outils plus récents de la génomique fonctionnelle, par exemple les mesures d'expression génique obtenue à l'aide des puces à ADN, ou encore les gels de protéines à deux dimensions, présentent encore des problèmes de fiabilité et de reproductibilité. Dans la mesure où les valeurs paramétriques, voire la forme même des fonctions impliquées sont mal connues, il est difficile d'établir la pertinence et la représentativité des résultats obtenus à l'aide de simulations quantitatives (continues ou stochastiques). En effet, pour des systèmes un tant soit peu complexes, les imprécisions sur les nombreux paramètres peuvent même présenter des effets multiplicatifs.

En réponse à ces difficultés, plusieurs groupes ont proposé des méthodes et des formalismes qualitatifs pour la modélisation des réseaux biologiques. Une première approche radicalise les propriétés de non-linéarité de ces systèmes en représentant la concentration de toute espèce moléculaire par une variable « booléenne » (ou « logique »), c'est-à-dire une variable qui ne peut prendre que deux valeurs, 0 ou 1 (ce qui peut être généralisé à plusieurs valeurs entières). On interprètera ces valeurs comme l'absence ou la présence de la molécule (protéine, co-facteur, etc.). L'évolution temporelle des valeurs des variables est définie par des équations logiques, où la valeur de chaque variable (présence ou absence d'une espèce moléculaire) dépend des valeurs précédentes d'une partie (voire de la totalité) des variables du système.

En fonction du type de traitement temporel, il faut distinguer les approches booléennes « synchrones » des approches « asynchrones » *(Figure 6)*. La première approche, la plus couramment utilisée en biologie théorique, impose que les valeurs de toutes les variables soient réactualisées simultanément à chaque itération *(Figure 6B)*. Au contraire, dans le cadre de l'approche asynchrone, on utilise des délais différents pour la réactualisation des différentes variables. Il va sans dire que

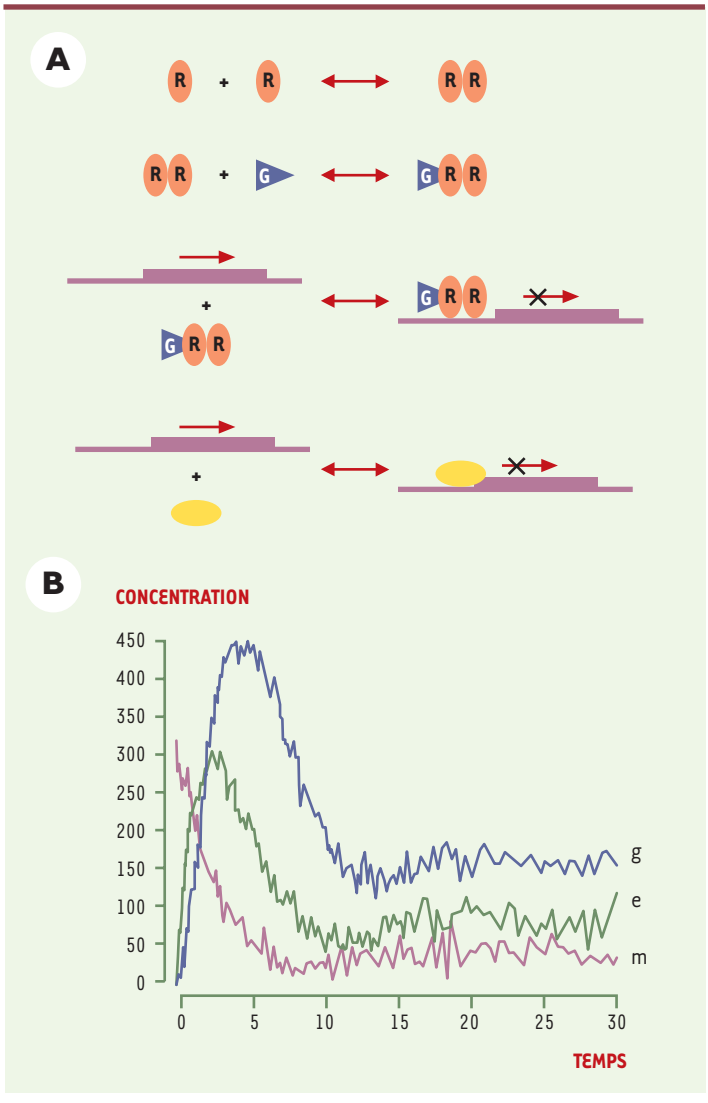


Figure 5. ***Modélisation stochastique du petit réseau d'inhibition d'une réaction par son produit*** (voir *Figure 4A*). ***A.*** Liste des réactions incluses dans le modèle. De haut en bas et dans le sens gauche-droite : dimérisation du régulateur R ; liaison du métabolite G au dimère R-R ; liaison du complexe répresseur G-R-R dans la région promotrice et répression de la transcription ; fixation de la polymérase et transcription en l'absence du complexe répresseur ***B.*** Résultat typique d'une simulation stochastique de ce réseau. Les trois courbes correspondent aux nombres de molécules d'ARNm (variable m, violet), de protéine enzymatique (variable e, vert), et du métabolite G (co-répresseur, variable g, bleu).

cette seconde approche, si elle est plus complexe à mettre en œuvre, apparaît beaucoup plus adéquate en biologie *(Figure 6C)*. En effet, il paraît abusif de considérer que les délais d'expression ou d'extinction des gènes aient tous la même valeur, même d'un point de vue qualitatif. Pratiquement, lors de simulations qualitatives, la simplification synchrone mène à des états de régime où plusieurs gènes changent indéfiniment de valeurs simultanément. Clairement artéfactuels, de tels cycles booléens synchrones ne correspondent à rien de similaire dans les modèles différentiels, stochastiques ou logiques asynchrones. Enfin, d'un point de vue biologique, il est important de réaliser que l'approche synchrone implique qu'à chaque état ne succède qu'un seul état au maximum, ce qui interdit toute représentation explicite des phénomènes de différenciation cellulaire.

Bien sûr, l'approche booléenne, même asynchrone, ne peut constituer qu'une représentation caricaturale des systèmes biologiques étudiés. Néanmoins, les modèles booléens permettent souvent de capturer les caractéristiques les plus marquantes du comportement dynamique des systèmes biologiques. En outre, pour des réseaux de taille raisonnable (de l'ordre de la dizaine d'éléments en interaction), il est possible de simuler de manière exhaustive les modèles booléens produits, vu le nombre peu élevé de valeurs permises pour chaque variable.

Au cours des dix dernières années, l'approche logique a été généralisée dans plusieurs directions (prise en considération de variables multi-valuées, introduction d'ensembles de paramètres logiques recouvrant plusieurs fonctions booléennes, représentation logique explicite des valeurs seuils séparant les valeurs entières, par exemple 0 et 1, etc) qui ont permis d'établir des correspondances étroites entre modélisations différentielle et logique [16, 29]. Mieux encore, ces généralisations conduisent à une approche beaucoup

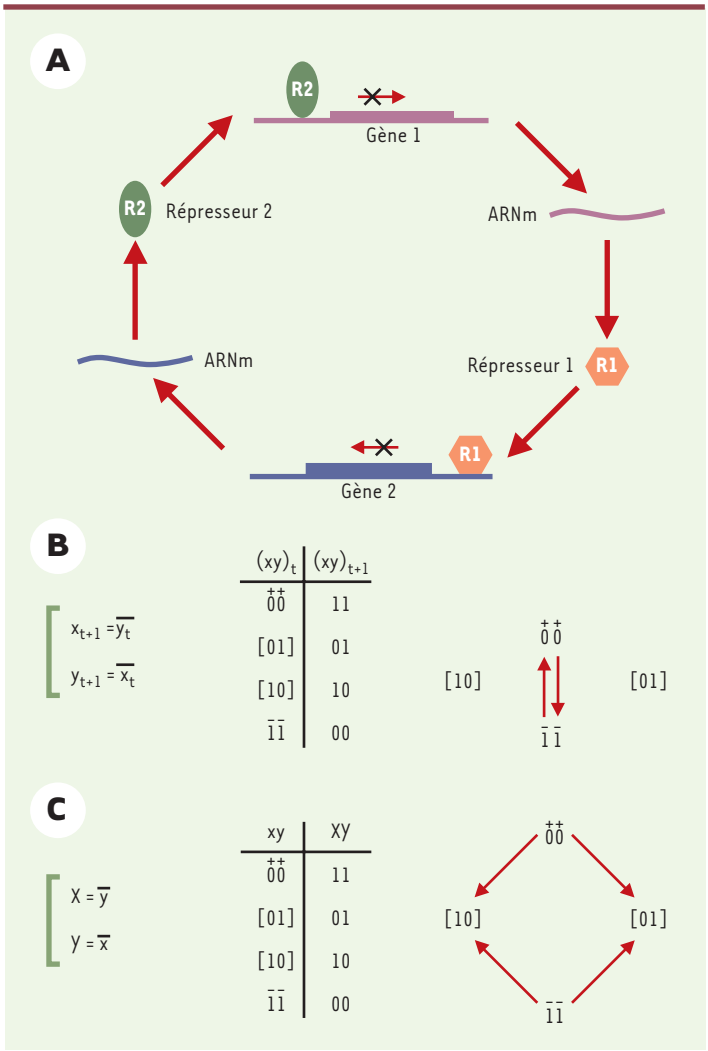


Figure 6. ***Illustration de la modélisation booléenne pour un système simple constitué de deux gènes s'inhibant mutuellement. A.*** Schéma moléculaire. ***B.*** Modélisation booléenne synchrone. Les équations (à gauche) expriment le fait qu'à un moment donné (t+1), le gène codant pour le premier répresseur (variable x) ne s'exprimera que si le second répresseur (variable y) était absent précédemment (à l'instant t) (idem pour l'expression d'y). La mise à jour des valeurs des deux variables se fait de manière simultanée, ce qui donne une table des états (milieu) et un graphe des séquences d'états (à droite) caractérisés par deux états stables avec expression exclusive de l'un ou l'autre des répresseurs. Le système comporte deux états stables, [10] et [01] (les crochets dénotant la stationnarité), où un seul répresseur est présent, inhibant la synthèse de l'autre de manière durable. Le système comporte un cycle où les deux gènes s'allument et s'éteignent de manière exactement synchrone et indéfiniment (flèches entre 00 et 11). En effet, à l'état 00, les deux répresseurs sont absents et les deux gène s'allument (« + » au-dessus des valeurs logiques 00). À l'instant suivant, les deux répresseurs seront considérés comme présents (état 11) et s'inhiberont l'un l'autre (« — » au-dessus des valeurs logiques 11), ramenant le système à l'état 00, etc. ***C.*** La formulation asynchrone ne prédétermine pas la manière de simuler le comportement temporel du système et représente son évolution sous la forme de fonctions « X » et « Y ». La table des états et le graphe des séquences d'états donnent les mêmes états stables que l'approche synchrone ([10] et [01]), mais les deux autres états conduisent maintenant vers l'un ou l'autre de ces états stables, en fonction des valeurs relatives des délais de transition associés à chacune des commutations (représentées par des flèches). D'un point de vue biologique, cela revient à dire qu'il y aura généralement un répresseur qui atteindra son seuil d'action plus rapidement que l'autre.

plus analytique de la dynamique des modèles logiques, en explicitant les liens existants entre certaines propriétés dynamiques (multistationnarité, comportement périodique) et la présence d'éléments topologiques particuliers dans les graphes d'interactions correspondants (circuits de régulation positifs ou négatifs). L'approche logique généralisée a été récemment appliquée à la modélisation de plusieurs réseaux de régulation génétique relativement complexe, en particulier le réseau contrôlant la différentiation des organes floraux chez la plante modèle *Arabidopsis thaliana* [30], ou encore des réseaux impliqués dans la formation de profils spatio-temporels au cours du développement embryonnaire de la drosophile [31, 32].

D'autres méthodes combinent une approche qualitative à des modèles utilisant des grandeurs variant de façon continue dans le temps. Les plus courantes décrivent les interactions entre espèces moléculaires en termes d'équations différentielles linéaires par morceaux [25], c'est-à-dire utilisant des équations linéaires définies pour chaque « boîte » encadrée par des valeurs de concentrations spécifiques (« seuils ») pour les différents composants du système. Les points de convergence entre cette description et la description logique généralisée sont nombreux. Ces méthodes ont été essentiellement utilisées dans le cadre de simulations, par exemple celle du réseau impliqué dans la détermination de la voie de sporulation chez la bactérie *Bacillus subtilis* [33]. Une telle approche devrait permettre un passage plus aisé du niveau qualitatif au niveau quantitatif.

D'autres méthodes hybrides sont actuellement développées, en particulier celles utilisant des « réseaux de Petri » [34]. Ces réseaux peuvent être considérés comme des graphes faisant intervenir deux types de sommets (digraphes): des « places » (espèces moléculaires) et des « transitions » (réactions). Les places peuvent contenir des « ressources » (molécules), susceptibles de circuler le long des arcs reliant les places *via* les transitions, en respectant les règles associées aux transitions. En partant d'un état initial donné, les règles de transitions permettent d'effectuer des simulations, déterministe ou stochastique suivant le cas. Les réseaux de Petri ont été récemment appliqués à des graphes métaboliques comportant des milliers de nœuds [35].

Développements récents et perspectives

Nous sommes donc actuellement en présence d'un foisonnement de méthodes de formalisation. Dans le domaine des réseaux moléculaires, les différentes approches qualitatives nous semblent offrir un cadre conceptuel proche de celui des biologistes moléculaires, et robuste vis-à-vis des incertitudes inhérentes aux données actuellement disponibles. Ces approches qualitatives permettent souvent de mettre en évidence les éléments et les interactions déterminants pour un réseau donné. Elles permettent également de prédire l'effet de perturbations drastiques d'un tel réseau (mutation perte de fonction, enclenchement d'un signal extracellulaire, etc.). Une fois les propriétés dynamiques et les conditions paramétriques correspondantes dégrossies, il est toujours possible de passer ensuite à une modélisation quantitative, éventuellement stochastique, afin de prédire de manière plus précise l'évolution des concentrations des espèces moléculaires impliquées, et ce pour différentes conditions. Au cours du processus de modélisation, il s'agit cependant avant tout de s'assurer d'une compréhension générale des propriétés du système étudié, plutôt que de s'attarder inconsidérément sur des simulations potentiellement anecdotiques. Par exemple, plusieurs auteurs ont mis en évidence la robustesse remarquable de plusieurs réseaux biologiques. Dans le cadre différentiel ou stochastique, cette robustesse se définit comme la capacité des systèmes correspondants de garder le même comportement dynamique qualitatif en dépit de modifications significatives des valeurs paramétriques [36].

À la suite du succès rencontré par les méthodes de mesure de l'expression génique à l'échelle génomique, la modélisation dynamique des réseaux génétiques suscite depuis quelques années un intérêt croissant. Plusieurs groupes se sont ainsi attelés à l'analyse des données de transcriptome dans le dessein d'inférer les réseaux d'interactions moléculaires sous-jacents (*reverse engineering*). À ce jour, ces efforts ont donné lieu à peu de résultats en raison : (1) du caractère incomplet et peu reproductible des données ; (2) de simplifications théoriques abusives ; et (3) de la sous-détermination des modèles du fait du nombre insuffisant de données expérimentales indépendantes. Si la dérivation de réseaux régulateurs *ab initio* sur la base des seules données d'expression paraît largement illusoire pour le moment, la combinaison d'outils d'inférence avec des données fonctionnelles diverses (fonctions connues ou prédites des gènes, données massives sur les interactions moléculaires protéine-protéine, ou protéine/ADN, etc.) devrait s'avérer très utile pour la caractérisation de réseaux de régulation impliqués dans des processus cellulaires spécifiques comme le cycle cellulaire, des voies de différenciation cellulaire, ou encore de leurs dérèglements pathologiques.

Récemment, les outils de modélisation et d'analyse dynamique ont servi de cadre théorique pour la synthèse

(➜) m/s 2001, n°1, p. 135

de petits réseaux génétiques à l'aide des outils du génie génétique [voir (➜) pour une revue de ces résultats]. Correspondant à de simples circuits de régulation positifs ou négatifs, ces constructions génétiques et leur analyse expérimentale ont permis de vérifier *in vivo* les propriétés dynamiques attribuées à ces deux classes de circuits (bistabilité dans le cas des circuits positifs, homéostasie ou expression périodique dans le cas des circuits négatifs). Plus remarquablement, l'état d'expression dynamique observé s'est chaque fois révélé transmissible d'une génération bactérienne à une autre, ainsi que relativement stable et robuste.

Afin de rendre plus accessible et plus systématique le travail de modélisation en biologie moléculaire, plusieurs groupes se sont attelés au développement de logiciels de modélisation, d'analyse et de simulation. Pour être vraiment utiles aux biologistes, de tels logiciels doivent être à la fois conviviaux, souples et performants. Il serait également utile de développer des articulations avec les bases de connaissances décrivant les macromolécules biologiques, leurs interactions, ou encore leur expression afin d'assister au mieux le biologiste dans le développement et la validation de modèles dynamiques. Une fois disponibles, de tels outils de modélisation dynamique devraient s'avérer rapidement incontournables pour caractériser les différents types de comportements dynamiques des réseaux biologiques, que ceux-ci soient normaux ou pathologiques, pour prédire les effets de perturbations environnementales ou imposées par l'expérimentateur ou le médecin, ou encore pour concevoir une nouvelle génération de constructions génétiques à même d'ajuster étroitement l'expression de gènes d'intérêt économique ou médical en fonction (ou en dépit) des variations du milieu intra- ou intercellulaire. ◊

SUMMARY

Modelling, analysis and simulation of gene networks

At the interface of biology, computer science and mathematics, a range of methods for the modelling, the analysis and the simulation of genetic regulatory networks have been developed. Such formal approaches allow the delineation of unambiguous descriptions of large and complex networks of interacting biological macromolecules, as well as predictions about their spatio-temporal behaviour, in normal or modified conditions. These theoretical methods can be subdivided into four broad categories depending on the mathematical formalism used: graphs, differential equations, stochastic models, and Boolean or generalised logical formalisms. Models further differ with respect to the level of granularity at which they describe regulatory networks, ranging from detailed molecular descriptions taking into account the stochastic effects arising from the small number of molecules of some components, to approximate models focusing on the global regulatory structure of a network. Several prokaryotic and eukaryotic regulatory networks have already been modelled and analysed, leading to new insights into the structure and functioning of these systems. Given the currently available genetic and molecular data, qualitative approaches, based on logical or differential equation models, seem particularly suitable. Still under development, these approaches and the corresponding computer tools should rapidly become indispensable for the study of the dynamical properties of normal and pathological biological networks, for the prediction of the effects of perturbations, and for the development of a new generation of therapeutic and agronomic tools. ◊

RÉFÉRENCES

1. Granjeaud S, Bertucci F, Jordan BR. Expression profiling: DNA arrays in many guises. *Bioessays* 1999 ; 21 : 781-90.
2. The Chipping Forecast. *Nat Genet* 1999 ; 21 (suppl janvier).
3. Iyer VR, Horak CE, Scafe CS, Botstein D, Snyder M, Brown PO. Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature* 2001 ; 409 : 533-8.
4. Ren B, Robert F, Wyrick JJ, *et al.* Genome-wide location and function of DNA binding proteins. *Science* 2000 ; 290 : 2306-9.
5. Pandey A, Mann M. Proteomics to study genes and genomes. *Nature* 2000 ; 405 : 837-46.
6. Tucker CL, Gera JF, Uetz P. Towards an understanding of complex protein networks. *Trends Cell Biol* 2001 ; 11 : 102-6.
7. Zhu H, Snyder M. Protein arrays and microarrays. *Curr Opin Chem Biol* 2001 ; 5 : 40-5.

8. Arnone MI, Davidson EH. The hardwiring of development: organization and function of genomic regulatory systems. *Development* 1997 ; 124 : 1851-64.
9. Hanahan D, Weinberg RA. The hallmarks of cancer. *Cell* 2000 ; 100 : 57-70.
10. Kohn KW. Molecular interaction map of the mammalian cell cycle control and DNA repair system. *Mol Biol Cell* 1999 ; 10 : 2703-34.
11. McAdams HH, Arkin A. Simulation of prokaryotic genetic circuits. *Annu Rev Biophys Biomol Struct* 1998 ; 27 : 199-224.
12. Smolen P, Baxter DA, Byrne JH. Modeling transcriptional control in gene networks - methods, recent results, and future directions. *Bull Math Biol* 2000 ; 62 : 247-92.
13. Van Helden J, Naim A, Mancuso R, *et al.* Representing and analysing molecular and cellular function using the computer. *J Biol Chem* 2000 ; 381 : 921-35.
14. Kanehisa M. Post-genome informatics. Oxford: Oxford University Press, 2000 : 148 p.
15. Wagner A, Fell DA. The small world inside large metabolic networks. *Proc R Soc Lond B Biol Sci* 2001 ; 268 : 1803-10.
16. Thomas R, Thieffry D, Kaufman M. Dynamical behaviour of biological regulatory networks. I. Biological role of feedback loops and practical use of the concept of the loop-characteristic state. *Bull Math Biol* 1995 ; 57 : 247-76.
17. Monod J, Jacob F. General conclusions: telenomic mechanisms in cellular metabolism, growth, and differentiation. *Cold Spring Harbor Symp Quant Biol* 1961 ; 26 : 389-401.
18. Thomas R, Thieffry D. Les boucles de rétroaction, rouages des réseaux de régulation biologiques. *Med Sci* 1995 ; 11 : 189-97.
19. Hlavacek WS, Savageau MA. Rules for coupled expression of regulator and effector genes in inducible circuits. *J Mol Biol* 1996 ; 255 : 121-39.
20. Tyson JJ, Novak B. Regulation of the eukaryotic cell cycle: molecular antagonism, hysteresis, and irreversible transitions. *J Theor Biol* 2001 ; 210 : 249-63.
21. Leloup JC, Goldbeter A. Modeling the molecular regulatory mechanism of circadian rhythms in Drosophila. *Bioessays* 2000 ; 22 : 84-93.
22. Meinhardt H, Gierer A. Pattern formation by local self-activation and lateral inhibition. *Bioessays* 2000 ; 22 : 753-60.
23. Reinitz R, Kosman D, Vanario-Alonso CE, Sharp D. Stripe forming architecture of the gap gene system. *Dev Genet* 1998 ; 23 : 11-27.
24. Von Dassow G, Meir E, Munro EM, Odell GM. The segment polarity network is a robust developmental module. *Nature* 2000 ; 406 : 188-92.
25. Edwards R, Siegelmann HT, Aziza K, Glass L. Symbolic dynamics and computation in model gene networks. *Chaos* 2001 ; 11 : 160-9.
26. Goodwin BC. Oscillatory behavior in enzymatic control processes. In Weber G, ed. *Advances in enzyme regulation*. Oxford: Pergamon Press, 1965 : 425-38.
27. Arkin A, Ross J, McAdams HH. Stochastic kinetic analysis of developmental pathway bifurcation in phage l-infected *Escherichia coli* cells. *Genetics* 1998 ; 149 : 1633-48.
28. Morton-Firth CJ, Shimizu TS, Bray D. A free-energy-based stochastic simulation of the Tar receptor complex. *J Mol Biol* 1999 ; 286 : 1059-74.
29. Thomas R. Regulatory networks seen as asynchronous automata: a logical description. *J Theor Biol* 1991 ; 153 : 1-23.
30. Mendoza L, Thieffry D, Alvarez-Buylla ER. Genetic control of flower morphogenesis in *Arabidopsis thaliana*: a logical analysis. *Bioinformatics* 1999 ; 15 : 593-606.
31. Sánchez L, van Helden J, Thieffry D. Establishment of the dorso-ventral pattern during the embryonic development of *Drosophila melagonaster*: a logical analysis. *J Theor Biol* 1997 ; 189 : 377-89.
32. Sánchez L, Thieffry D. A logical analysis of the gap gene system. *J Theor Biol* 2001 ; 211 : 115-41.
33. De Jong H, Page M, Hernandez C, Geiselmann J. Qualitative simulation of genetic networks: method and application. In: Nebel B, ed. *Proceedings of the 17th Int Joint Conf Artif Intell 2001 - IJCAI-01*. Morgan Kaufmann, 2001: 67-73.
34. Goss PJ, Peccoud J. Quantitative modeling of stochastic systems in molecular biology by using stochastic Petri nets. *Proc Natl Acad Sci USA* 1998 ; 95 : 6750-5.
35. Kuffner R, Zimmer R, Lengauer T. Pathway analysis in metabolic databases *via* differential metabolic display (DMD). *Bioinformatics* 2000 ; 16 : 825-36.
36. Barkai N, Leibler S. Robustness in simple biochemical networks. *Nature* 1997 ; 387 : 913-7.

TIRÉS À PART

D. Thieffry