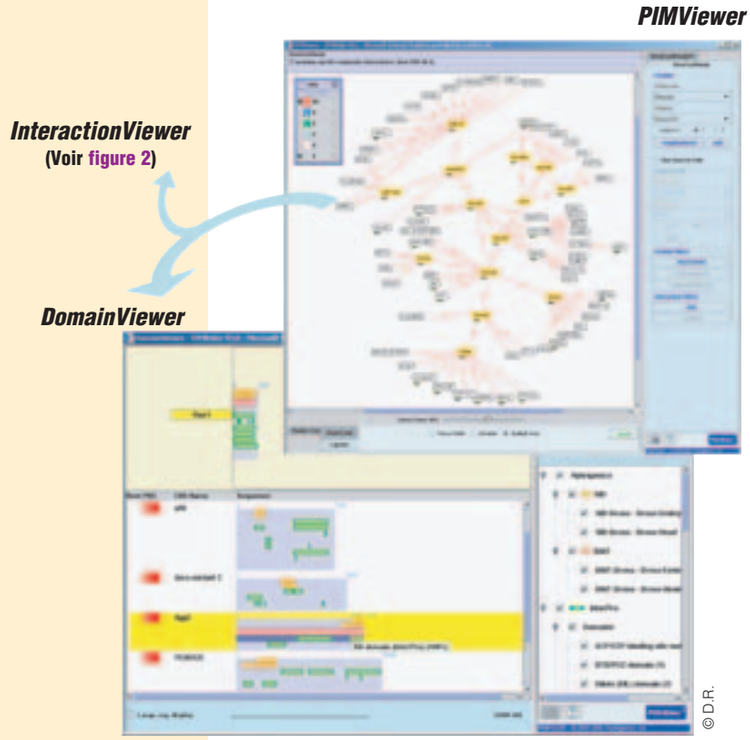


Figure 4 ***PIMRider*, plate-forme d'exploration des cartes d'interactions protéine-protéine.**
Parmi les applets constitutives de cette plate-forme, le *PIMViewer* permet de naviguer et d'explorer tous les partenaires d'un réseau, ici celui des protéines de la voie de signalisation Smad, éventuellement éclairci sur la base des valeurs du score de confiance **(figure 3)**. Le *DomainViewer* se charge d'offrir au chercheur les moyens de visualiser les domaines constitutifs des partenaires d'interaction, en particulier les domaines InterPro figurant ici en vert. Indissociable, un troisième outil (l'*InteractionViewer*, voir **figure 2**) se consacre aux particularités et détails de chaque interaction.

de cette information et son enrichissement, jusqu'à son exploitation *via* des logiciels dédiés. Cet outil est indispensable pour garantir la qualité de l'information, et notamment réduire le taux de faux-positifs lié aux limitations expérimentales et aux contraintes de haut débit. Grâce à ce fondement bioinformatique robuste, l'information contenue dans les réseaux d'interactions protéiques peut ensuite être validée dans des systèmes cellulaires adéquats, permettre une meilleure compréhension des mécanismes d'action moléculaire et, *in fine*, proposer des modes d'action thérapeutique. ●

Modélisation, analyse et simulation de réseaux de régulation génique

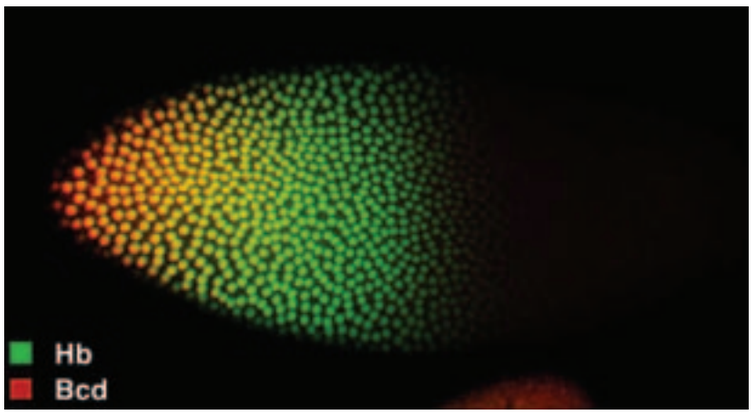
L'analyse des réseaux de régulation génique contrôlant le fonctionnement et le développement des organismes vivants nécessite des techniques expérimentales performantes, mais aussi des outils de modélisation, d'analyse et de simulation informatiques. Nous présentons les principales approches existantes, accompagnées de quelques exemples d'applications.

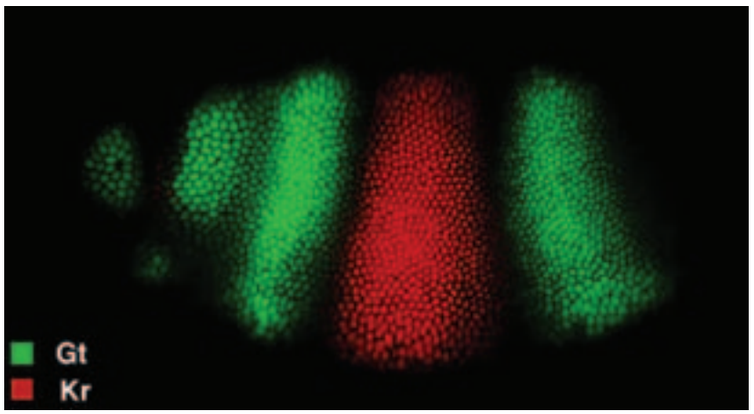
Hidde de Jong*, Delphine Ropers*, Claudine Chaouiya, Denis Thieffry****

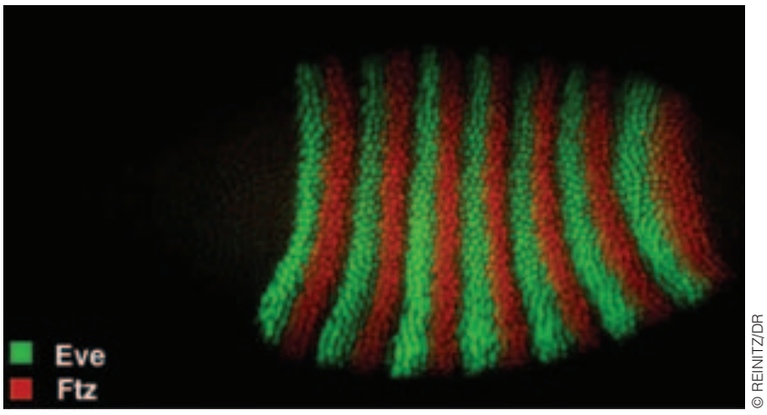
* Inria Rhône-Alpes, 655 avenue de l'Europe, Montbonnot, 38334 St Ismier Cedex ; Hidde.de-Jong@inrialpes.fr, Delphine.Ropers@inrialpes.fr

** LGPD-IBDM, Campus de Luminy, CNRS Case 907, 13288 Marseille Cedex 9 ; thieffry@ibdm.univ-mrs.fr, chaouiya@esil.univ-mrs.fr

Les développements récents en génomique ouvrent de nouvelles perspectives pour le décryptage des mécanismes de régulation contrôlant le développement, la croissance et les processus physiologiques chez les êtres vivants. Agissant à divers niveaux (transcription et traduction du matériel génétique, modifications des protéines, etc.) et souvent de manière croisée, ces mécanismes régulateurs forment des *réseaux complexes*, qui constituent vraisemblablement le niveau fonctionnel le plus pertinent pour comprendre quand, où et comment leurs composants sont activés. La modélisation, l'analyse et la simulation de ces réseaux sont ainsi de plus en plus considérées par les biologistes comme un élément essentiel au sein de leur panoplie analytique.







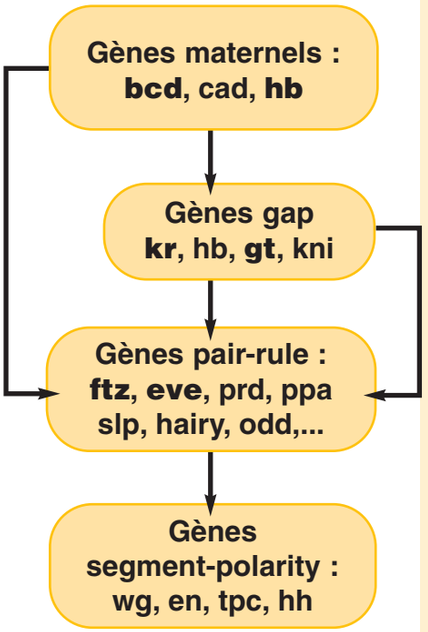


Figure 2 **Modélisation logique du réseau contrôlant la segmentation au cours du développement précoce de la mouche *Drosophila melanogaster* (9).** À gauche : patrons d'expression pour six facteurs de régulation, deux par image, révélés par anticorps *in situ*. Chaque tâche lumineuse correspond à un noyau (ces images ont été fournies par John Reinitz (SUNY at Stony Brook, NY, États-Unis). À droite : principales classes de gènes impliqués dans le contrôle de la segmentation, et représentation schématiques des relations de régulation entre ces classes (des patrons d'expression pour les gènes en gras sont proposés dans la partie de gauche).

Pourquoi modéliser ?

Afin d'intégrer des données expérimentales nombreuses et variées et de comprendre comment le comportement dynamique d'une cellule ou d'un organisme émerge des interactions entre ses composants, nous avons besoin d'outils *informatiques de modélisation, d'analyse et de simulation*. D'une part, ces outils permettent d'organiser les connaissances disponibles sur un réseau de régulation sous la forme d'un modèle mathématique ou informatique. D'autre part, à partir de cette représentation formelle, il devient possible d'utiliser l'ordinateur afin de prédire le comportement du système pour différentes conditions physiologiques ou environnementales. De telles simulations contribuent non seulement à une meilleure compréhension du rôle des différents composants moléculaires et interactions d'un processus cellulaire donné, mais peuvent également révéler des implications inattendues du modèle, susceptibles de suggérer de nouvelles expériences. Alliant les techniques expérimentales issues de la génomique et les outils de modélisation, cette démarche s'inscrit dans un domaine connu sous des noms divers tels que «biologie des systèmes», «biologie intégrative» ou «génomique fonctionnelle» (1).

Comment modéliser ?

Différentes approches peuvent être utilisées (pour des revues et de plus amples références à la littérature, voir (2-4)). Chacune d'elles s'appuie sur un formalisme mathématique ou informatique particulier, qui permet de représenter un réseau sous forme d'un modèle. Nous présentons ici brièvement quatre exemples de formalisme : les *graphes*, les *réseaux logiques*, les *équations différentielles* et les *équations stochastiques*.

Le choix de l'un de ces formalismes dépend avant tout des questions biologiques que l'on souhaite étudier. Comme nous le préciserons plus loin, chaque formalisme se focalise sur certaines propriétés du réseau et en néglige d'autres. En outre, la disponibilité de données biologiques impose des contraintes sur le choix de la formalisation, car le degré de précision de la description des réseaux – et donc le besoin d'informations sur les facteurs de régulation et leurs interactions – varie d'un formalisme à un autre. Dans les paragraphes qui suivent, nous introduisons ces différentes approches de modélisation à travers plusieurs applications, portant toutes sur des réseaux de régulation génique.

(1) H Kitano (2002) *Science* 295, 1662-4
(2) H Bolouri, EH Davidson (2004) *Bioessays* 24, 1118-29
(3) H de Jong (2002) *J. Comput. Biol.* 9, 67-103
(4) D Thieffry, H de Jong (2002) *Médecine Science* 18, 492-502

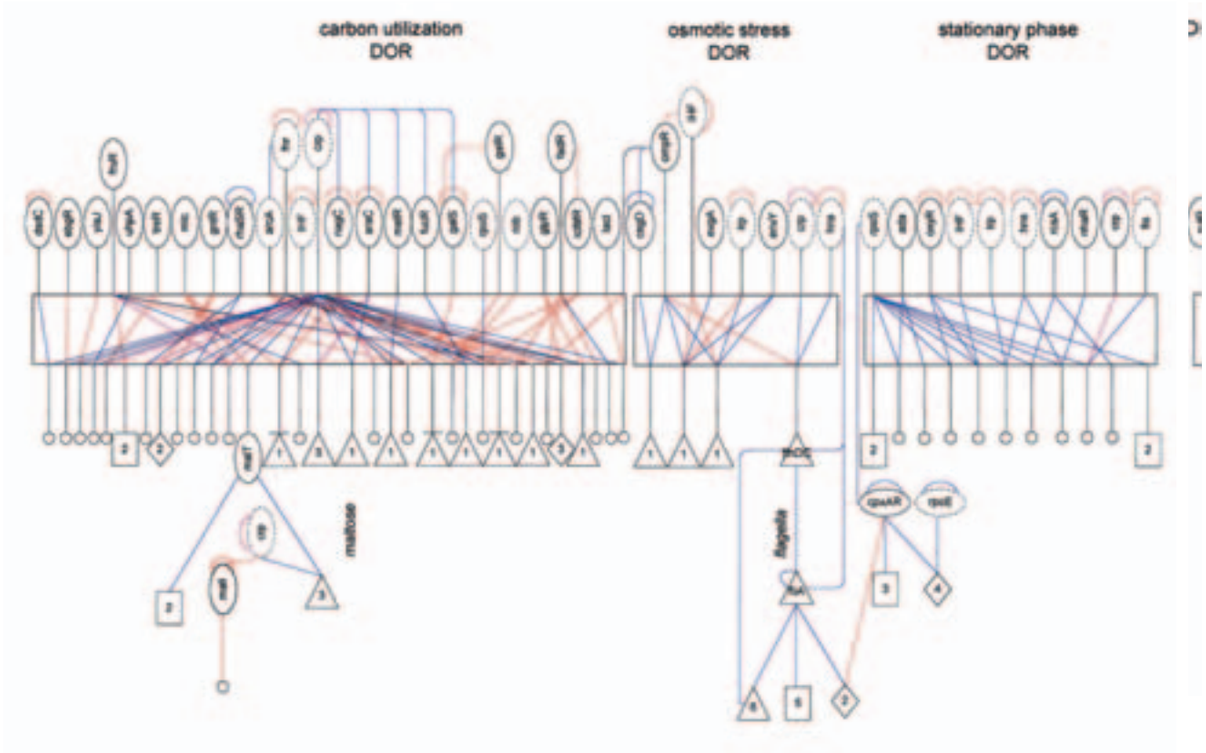


Figure 1 **Partie du réseau des régulations transcriptionnelles connues chez *Escherichia coli*,** représentée sous forme d'un graphe **(7)**.

(5) KW Kohn (1999) *Mol. Biol. Cell* 10, 2703-34
(6) D Thieffry *et al.* (1998) *Bioessays* 20, 433-40
(7) SS Shen-Orr *et al.* (2002) *Nat. Genet.* 31, 64-8

Graphes d'interactions ou de régulation

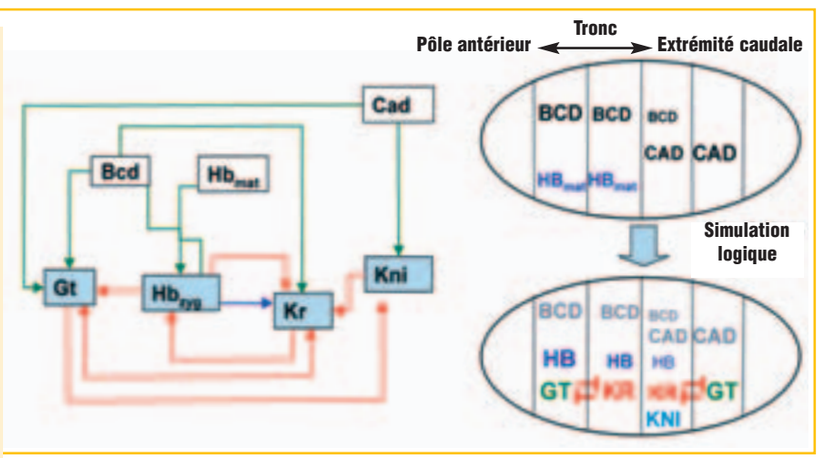
Pour le biologiste, l'utilisation de schémas constitue généralement la manière la plus naturelle de représenter les réseaux de régulation (voir par exemple les travaux de K.W. Kohn **(5)**). Formellement, de tels schémas peuvent se traduire de façon rigoureuse et homogène en termes de *graphes*, où les nœuds représentent par exemple des gènes et les arêtes, des interactions régulatrices. Pour des organismes bien caractérisés, comme la levure ou le colibacille, il est même possible d'extraire les informations encodées dans des bases de données publiques, afin de générer des graphes représentant les réseaux métaboliques ou génétiques à une échelle génomique.

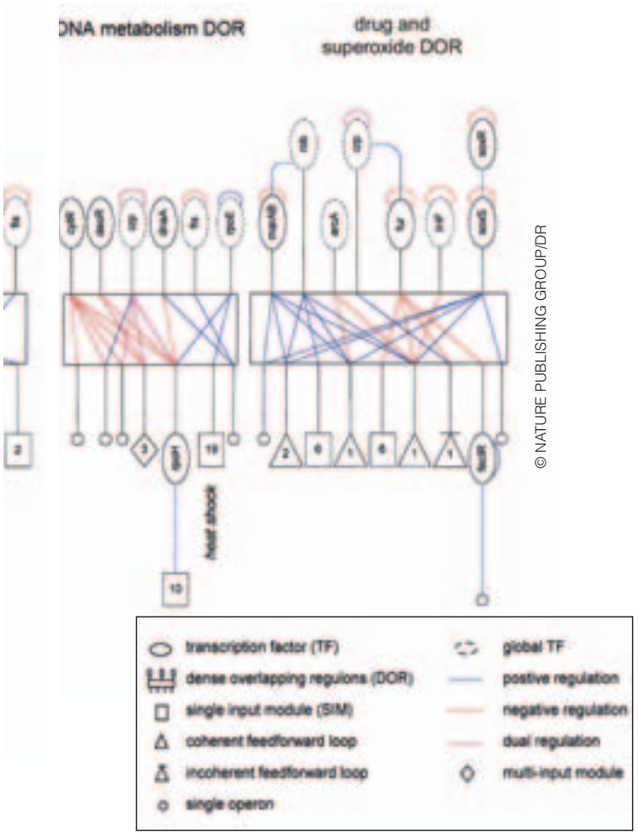
L'analyse de tels graphes a déjà permis la mise en évidence de motifs interactifs beaucoup plus fréquents que ce que l'on attendrait dans le cas de graphes avec des connexions distribuées aléatoirement. Par exemple, dans le graphe rassemblant les régulations transcriptionnelles connues chez *Escherichia coli* **(figure 1, ci-dessus)**, on trouve un très grand nombre d'autorégulations négatives, ce qui suggère une régulation homéostatique de la concentration des facteurs de transcription correspondants **(6)**. D'autre part, des études plus récentes ont permis la mise en évidence de motifs plus complexes, néanmoins très fréquents, comme les motifs «*feed forward*», où un gène régule un autre gène directement, ainsi qu'indirectement, *via* la régulation d'un troisième gène **(7)**.

Modèles logiques

Si l'analyse des graphes biologiques apporte des informations intéressantes concernant leurs principes d'organisation, il faut cependant recourir à des méthodes de modélisation dynamique pour apprécier pleinement l'impact fonctionnel de tels principes, en particulier lorsque les graphes se compliquent quelque peu. On

Figure 3 **Modélisation logique du réseau contrôlant la segmentation au cours du développement précoce de la mouche *Drosophila melanogaster* (9)**
À gauche : régulations croisées entre gènes *gap* et régulations des gènes *gap* par les principaux gènes maternels ; les flèches en vert, rouge et bleu correspondent respectivement à des activations, inhibition et à une interaction duale (positive ou négative suivant le contexte). Bcd et Hb régulent de manière synergique Hb. À droite : simulation logique pour le cas sauvage. À partir d'un état initial donné par des gradients de concentration des produits maternels, le système atteint un état caractérisé par une expression différentiée des gènes *gap,* suggérée par des variations de taille de caractères pour les noms correspondants. Le circuit positif formé des inhibitions croisées entre Gt et Hb est à l'origine de leurs patrons d'expressions exclusifs.





peut alors utiliser un *formalisme logique* pour définir les effets des combinaisons d'interactions (régulations) arrivant sur chaque nœud (gène). Le modèle logique le plus simple utilise des variables et fonctions booléennes, qui ne peuvent prendre que deux valeurs, 0 ou 1, selon qu'un régulateur est absent ou présent, ou qu'un gène est éteint ou allumé. Le recours à des opérateurs logiques permet alors d'exprimer assez simplement, par exemple, le fait que plusieurs activateurs doivent être présents pour permettre l'expression d'un gène (ET logique), ou qu'un ou plusieurs répresseur(s) doivent au contraire être absents (NON logique). En combinant de tels opérateurs, on peut définir des fonctions de régulation très complexes. Dans le cadre d'une telle approche qualitative, on peut simuler le comportement du système, en partant d'un état initial représentant une configuration de composants moléculaires présents ou absents, et en calculant les états suivants, de manière itérative, en utilisant les fonctions logiques associées à chaque élément (gène).

L'approche logique peut être affinée, en prenant en compte plusieurs niveaux d'expression qualitativement différents pour certains gènes, ou encore en tenant compte des variations dans les délais de synthèse des régulateurs **(8)**. Le potentiel d'une telle approche a été démontré dans le cadre de la modélisation du réseau de régulation génique contrôlant le processus de segmentation au cours du développement de l'embryon de la drosophile **(figures 2, page 37 et 3 ci-contre)**. L'approche logique permet ainsi la reconstitution qualitative des principales étapes de différenciation caractérisées expérimentalement, ainsi que la prédiction du phénotype pour diverses mutations ou autres perturbations. En outre, elle apporte des informations sur les rôles dynamiques et fonctionnels de structures particulières dans le graphe de régulation, en particulier celui de circuits positifs, toujours à l'origine des décisions différenciatives (**(9)** ; une revue comparant différents modèles pour ce système est proposée par D.Thieffry et L. Sánchez **(10)**).

Équations différentielles

Le formalisme le plus utilisé pour modéliser les réseaux de régulation génique est sans aucun doute celui des *équations différentielles*. Suivant cette approche, les concentrations des composants moléculaires du système – les protéines, les ARNm, les petites molécules de signalisation – sont représentées par des variables réelles positives qui évoluent de manière continue au cours du temps. Le taux de variation de la concentration d'un composant moléculaire dépend évidemment de la concentration des autres composants, à travers des interactions régulatrices. Les équations différentielles décrivent de telles interactions en exprimant la dérivée temporelle d'une variable (concentration) en fonction de la valeur des autres variables. Pour résoudre les équations différentielles, c'est-à-dire préciser comment les concentrations des différents composants moléculaires évoluent au cours du temps, on dispose d'une grande variété de techniques d'analyse et de simulation qui s'appuient sur des outils informatiques performants.

L'application de ces techniques a été particulièrement bien illustrée dans le cadre de la modélisation du réseau d'interactions contrôlant la division cellulaire chez les eucaryotes **(figure 4)**. Sur la base de nombreux travaux expérimentaux, les groupes de B. Novak et de J. Tyson ont construit des modèles d'équations différentielles qui incluent les interactions contrôlant l'activité des «*cyclin-dependent kinases*» (CDK), protéines dont dépendent les événements majeurs du cycle cellulaire **(11)**. L'analyse et la simulation de ces modèles, comprenant plusieurs dizaines d'équations, ont produit des prédictions testables du phénotype de la cellule dans le cas sauvage, ainsi que pour un grand nombre de mutations. Récemment, l'une de ces prédictions a été confirmée expérimentalement, alors que d'autres ont été contredites par les observations, suggérant la nécessité de prendre en compte des interactions jusque-là ignorées **(12)**.

La plupart des techniques d'analyse et de simulation différentielles requièrent une détermination précise des paramètres cinétiques et des concentrations moléculaires, alors que ces valeurs ne sont que rarement disponibles. Face à ce problème, certains groupes ont développé des approches numériques en se basant sur l'hypothèse que les propriétés essentielles du système sont robustes face aux variations de valeurs paramétriques **(13)**. D'autres chercheurs ont développé des techniques d'analyse qui ne dépendent pas de valeurs paramétriques exactes, en mettant donc l'accent sur des propriétés qualitatives **(14)**. Cette approche, appliquée à l'analyse de réseaux de régulation bactériens **(15)**, rappelle les intuitions sous-jacentes des modèles logiques, mais les développe dans le cadre formel des équations différentielles.

Modélisation stochastique

Une critique courante concernant le recours aux équations différentielles est que ces dernières font abstraction de la nature stochastique des réactions biochimiques. Les fluctuations aléatoires au niveau moléculaire impliquent que deux systèmes ayant les mêmes conditions

(8) R Thomas (1991) *J. theor. Biol.* 153, 1-23
(9) L Sánchez, D Thieffry (2003) *J. theor. Biol.* 224, 517-37
(10) D Thieffry, L Sánchez (2003) *Curr. Opin. Genet. Dev.* 13, 326-30
(11) JJ Tyson *et al.* (2001) *Nat. Rev. Mol. Cell Biol.* 2, 908-16
(12) FR Cross *et al.* (2002) *Mol. Biol. Cell* 13, 52-70
(13) G von Dassow *et al.* (2000) *Nature* 406, 188-92
(14) H de Jong *et al.* (2004) *Bull. Math. Biol.* 66, 301-40
(15) H de Jong *et al.* (2004). *Bull. Math. Biol.* 66, 261-300

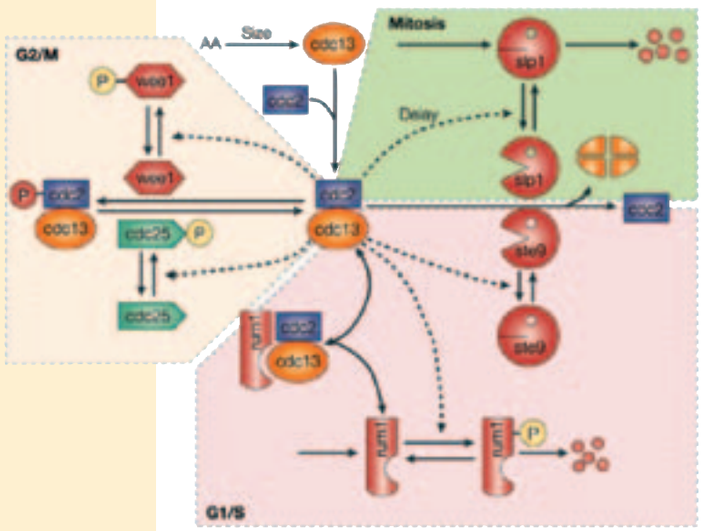


Figure 4 **Réseau moléculaire** contrôlant le cycle cellulaire chez les eucaryotes, modélisé par des équations différentielles **(11)**.

initiales peuvent éventuellement atteindre des états différents, en particulier lorsque le nombre de molécules de chaque espèce dans la cellule est petit. Afin de prendre en compte les aspects stochastiques, les modélisateurs ont recours à d'autres formalismes. Une première approche consiste à inclure un terme de bruit dans les équations différentielles (*équations différentielles stochastiques*), tandis qu'une deuxième approche utilise les équations maîtresses stochastiques, bien connues en physique. Ces équations décrivent l'évolution de la probabilité des différents états possibles du système, caractérisés par le nombre de molécules de chaque espèce présent dans la cellule. Pour résoudre ces équations, des méthodes de simulation stochastique ont été développées, souvent très gourmandes en calcul à cause du nombre élevé de molécules et de réactions impliquées.

Les modèles stochastiques ont été utilisés pour explorer l'origine de la variation de l'expression génique d'une cellule à une autre dans une population génétiquement identique **(16)**. Les prédictions du modèle, qui tiennent l'efficacité de la traduction (plutôt que celle de la transcription) pour responsable de la variation de l'expression génique entre cellules, ont été confirmées expérimentalement, à l'aide de mesures d'expression d'un gène rapporteur codant une protéine fluorescente. Comme A. Arkin et ses collègues l'ont souligné dans le cas de la réponse d'*Escherichia coli* à l'infection par le bactériophage lambda, les fluctuations du niveau de concentration des protéines peuvent avoir des conséquences importantes pour le phénotype de la cellule **(17)**. Un inventaire minutieux des réactions entre les gènes et protéines-clés du bactériophage a conduit à l'élaboration d'un modèle qui explique comment, suite à l'infection par le phage, une population de bactéries, initialement homogène, se répartit en deux sous-populations phénotypiquement distinctes. Les simulations stochastiques menant à ces conclusions sont difficiles à réaliser pour d'autres systèmes, parce qu'elles demandent des connaissances précises sur les mécanismes de réaction et les paramètres cinétiques. Heureusement, les simulations stochastiques sont souvent en accord avec les prédictions obtenues à partir de modèles déterministes, en partie suite aux effets de rétroactions négatives ou d'autres caractéristiques topologiques du réseau qui favorisent la robustesse du comportement de la cellule en dépit des fluctuations moléculaires **(18)**.

Conclusions : réalisations et défis actuels

Le biologiste a donc à sa disposition un grand nombre de formalismes et d'outils pour modéliser, analyser et simuler les réseaux de régulation complexes auxquels il est de plus en plus confronté. Comme les différents exemples cités plus haut le suggèrent, la modélisation dynamique peut apporter des éléments de compréhension nouveaux et des résultats concrets sous forme de prédictions expérimentales dans de nombreux cas de figure, que le système étudié soit déjà bien caractérisé ou non. En revanche, il s'agit de bien choisir le formalisme de modélisation (qualitatif *versus* quantitatif) en fonction des données disponibles, des questions posées, ainsi que des approches expérimentales disponibles pour le système étudié. À tout le moins, la démarche de modélisation conduit d'emblée à préciser et structurer les connaissances disponibles.

La maîtrise des outils de modélisation n'est pas encore très répandue parmi les biologistes, et il s'agira donc le plus souvent d'établir des collaborations interdisciplinaires pour pouvoir pleinement exploiter leur potentiel dans le cadre d'applications sophistiquées. Les exemples de projets de recherche combinant harmonieusement expérimentation et modélisation restent ainsi très rares. Les différences de culture scientifique entre biologistes, mathématiciens et informaticiens compliquent les échanges et les collaborations, d'autant plus qu'il s'agit en fait souvent de faire progresser simultanément les problématiques biologiques et les méthodes formelles. Ceci est bien illustré par la problématique de l'intégration de nos connaissances sur les réseaux métaboliques, les réseaux génétiques et les réseaux de transduction de signaux **(19)**. Ces différents réseaux font intervenir des réactions ou régulations correspondant à des échelles de temps variées, des migrations entre compartiments intracellulaires, ainsi que des interactions entre cellules à l'intérieur de l'organisme. Pour modéliser des systèmes aussi complexes, nous avons besoin de méthodes permettant une approche modulaire de la modélisation, de manière à pouvoir combiner des modèles de processus spécifiques (cycle cellulaire, voies de transduction de signaux, petits réseaux transcriptionnels, etc.), pour constituer des représentations formelles de cellules, tissus, organes, voire d'organismes entiers. Il s'agit donc en quelque sorte de développer des modèles emboîtés, susceptibles de décrire les différents niveaux de régulation impliqués dans les processus vivants de manière cohérente, pour faciliter le passage d'une échelle à une autre. Dans le même esprit, il est important d'établir des passerelles entre les méthodes de modélisation qui correspondent à des niveaux de précision différents, depuis les descriptions graphiques ou logiques aux équations différentielles ou stochastiques. ●

(16) EM Ozbudak *et al.* (2002) *Nat. Genet.* 31, 69-73
(17) A Arkin *et al.* (1998) *Genetics* 149, 1633-48
(18) D Gonze *et al.* (2002) *Proc. Natl. Acad. Sci. USA* 99, 673-8
(19) MW Covert *et al.* (2004) *Nature* 429, 92-6