Méthodes informatiques en biologie

Georgia Barlovatz-Meimon[*,+] et Sylvain Sené[**,§]

[*] Université d'Évry – Val d'Essonne, IBISC, EA 4625, 91000 Évry, France
[+] Université Paris-Est-Créteil, 94000 Créteil, France
[**] Aix-Marseille Université, CNRS, LIF, UMR7279, 13000 Marseille, France
[§] Institut rhône-alpin des systèmes complexes, IXXI, 69000 Lyon, France

1. Introduction

L'informatique, qu'elle soit considérée comme une science ou une technologie, tient une place croissante dans le développement des recherches en biologie. Il suffit pour s'en convaincre de considérer le grand nombre de revues, de conférences et plus généralement de publications à la frontière de l'informatique et de la biologie depuis une quinzaine d'années[1]. Au delà de ce simple constat, peut-être est-il intéressant de comprendre et d'expliquer pourquoi. Pour ce faire, il convient de se plonger dans l'histoire de ces deux disciplines scientifiques et d'en trouver les points de liaison caractéristiques. Du point de vue de la biologie, l'on considère assez généralement que les premières utilisations remarquables de l'informatique dans le domaine se sont produites à partir des années 1970, au moment des premiers pas faits parallèlement par Sanger et Gilbert vers le séquençage de génomes à A.D.N. (Maxam, Gilbert, 1977 ; Sanger, Air et al., 1977), qui ont soulevé la nécessité de développer des méthodes informatiques pour traiter la quantité de données émanant des génomes entiers. Sans développer plus avant les méthodes en question dès maintenant, malgré les recherches importantes menées en informatique dans ce contexte, cette vision d'un tel apport de l'informatique à la biologie se positionne principalement au travers de la dimension technologique et appliquée de la discipline informatique. Or, l'histoire montre que les premières intersections de ces deux disciplines sont plus anciennes et mettent en jeu la science informatique dans ses aspects fondamentaux. La première étude notable ayant ainsi combiné les connaissances accumulées dans ces deux disciplines date en réalité des années 1940 et a été menée par McCulloch et Pitts (McCulloch, Pitts, 1943). En proposant une première modélisation des interactions neuronales sous forme de réseaux, elle s'est construite autour de la volonté de comprendre les capacités de calcul induites par ces réseaux en considérant les neurones comme les unités logiques de base des calculs.

Les deux exemples évoqués, autour du séquençage des génomes et des réseaux de neurones, diffèrent significativement dans les questionnements qu'ils soulèvent et les réponses qu'ils apportent aussi bien en biologie qu'en informatique. Là où la première est de caractère pluridisciplinaire (les développements de l'informatique apportent à la biologie), la seconde est de nature interdisciplinaire (les deux disciplines œuvrent de concert et les développements de l'une participent aux développements de l'autre, et vice versa). Cela nous amène naturellement à introduire le terme de *bio-informatique*, dont le champ sémantique est large et qu'il convient d'appréhender en connaissance de cause. À l'origine, ce terme a été défini par Hesper et Hogeweg (Hogeweg, Hesper, 1978) comme l'étude des processus informatiques dans les systèmes biotiques. Cette définition donne la part belle à l'informatique et limite le rôle de la biologie à celui de domaine d'applications, ce qui peut surprendre à première vue de la part de biologistes de formation. Aujourd'hui, l'acception commune de la bio-informatique est autre et désigne souvent, de manière trop simpliste,

1 Cette croissance est également largement visible du point de vue économique. À titre d'exemple, le couplage de la biologie et de l'informatique a représenté un marché global estimé à 2,3 milliards de dollars en 2012, prévu pour atteindre 9,1 milliards de dollars en 2018 (Transparency Market Research, 2012).

l'utilisation des ordinateurs en biologie[2]. Par conséquent, du point de vue scientifique, il convient de sortir de cette vision réductrice. Dans ce contexte, la bio-informatique peut se distinguer selon deux aspects :

- le premier, le plus classique, aborde des problèmes précis posés par la biologie, sur la base de données relativement fiables issues des expérimentations biologiques. On trouve dans ces problèmes ceux ayant trait à l'identification automatique de séquences d'A.D.N., aux études de la structure des protéines, aux bases de données regroupant par exemple les annotations, à la phylogénie. Ce type de bio-informatique développe des méthodes informatiques pour y répondre. Les domaines de l'informatique largement utilisés sont, entre autres, l'algorithmique, l'imagerie, les bases de données... Les solutions apportées sont généralement utilisables en biologie à court terme ;
- le second, qui se rapproche de la définition originale de la bio-informatique, est souvent appelé différemment (biologie des systèmes, modélisation des systèmes biologiques complexes...). Son objectif principal est de dégager les lois générales qui régissent le vivant, en étudiant les grandes lignes des échanges d'information et de leurs conséquences au sein de processus biologiques à large échelle comme les régulations, la morphogenèse ou encore la croissance des organismes, décrits par des données peu fiables, et parfois absentes. Les méthodes informatiques utilisent alors des abstractions plus élevées, comme celles émanant des modèles de calculs et des systèmes dynamiques discrets. Ici, les problèmes caractéristiques mettent en jeu des mécanismes de traitement et de transmission de l'information intrinsèquement complexes pertinents dans les deux disciplines.

La combinaison de ces deux aspects permet d'accroître significativement la compréhension des mécanismes biologiques, tout en utilisant les problématiques inhérentes à la biologie afin de développer les questions et les méthodes qui leur sont propres. À titre d'exemple, les recherches menées sur la découverte automatique de séquences d'A.D.N. permettent à celles tournées vers la modélisation des réseaux de régulation génétique d'apporter leur lot de réponses, parfois éloignées de la réalité, approximant toujours cette dernière mais néanmoins toujours utiles pour augmenter la connaissance quant aux lois biologiques gouvernant les processus de morphogenèse.

L'objectif de ce chapitre est dans un premier temps de donner un aperçu des grands thèmes qui constituent la bio-informatique contemporaine, en précisant notamment les techniques informatiques qui s'en trouvent au centre. Ensuite, nous nous attacherons à présenter un domaine de la bio-informatique qui nous semble parmi les plus originaux et les plus porteurs, celui de la modélisation des réseaux de régulation. Plus précisément, nous livrerons nos points de vue qui sont ceux d'une biologiste cellulaire et d'un informaticien, ce qui permettra au lecteur de considérer et d'appréhender les divergences d'opinion, mais nous verrons surtout que, malgré ces différences (qui proviennent essentiellement du fait que là où l'informatique se pose essentiellement des questions sur les méthodes, la biologie s'attache principalement aux objets), ces deux disciplines œuvrent de concert pour contribuer à la connaissance scientifique, qu'elle soit appliquée ou théorique.

2. Vision globale de la bio-informatique actuelle

2.1 Bio-informatique des séquences

Qu'il s'agisse de séquences de nucléotides ou de séquences d'acides aminés, il s'agit toujours d'enchainements d'unités élémentaires. Pour l'A.D.N., il s'agit des quatre bases, A., C., G., T.. Pour l'A.R.N.,

2 Cette acception est effectivement simpliste dans la mesure où elle réduit l'informatique à l'ordinateur alors que l'ordinateur n'est qu'un objet technologique, né en 1946 des développements de la recherche en science informatique, ou science du calcul, dont les premiers travaux peuvent être ramenés à ceux de Gödel (Feferman, Dawson et al., 1986), Shannon (Sloane, Wyner, 1993), Church (Church, 1941) et bien sûr Turing (Turing, 1936), puisant eux même leurs sources dans les travaux de Fredge et Russel sur la logique (Fredge, 1884 ; Whitehead, Russel, 1910) et d'Ada Lovelace (Lovelace, Toole, 1992) sur les algorithmes.

ce sont les bases A., C., G., U. et, pour les protéines, les vingt acides aminés, Ala (A.), Cys (C.), Asp (D.), Glu (E.), Phe (F.), Gly (G.), His (H.), Ile (I.), Lys (K.), Leu (L.), Met (M.), Asn (N.), Pro (P.), Gln (Q.), Arg (R.), Ser (S.), Thr (T.), Val (V.), Trp (W.), Tyr (Y.). Les chaînes de nucléotides et d'acides aminés peuvent être représentées par une succession ordonnée et orientée d'unités élémentaires identifiées.

L'un des premiers aspects de la bio-informatique des séquences est celui qui traite du séquençage lui-même, c'est-à-dire connaître l'enchaînement complet des bases nucléotidiques qui constituent un génome. Si le séquençage d'un gène ne nécessite aucune méthode propre à la bio-informatique (Maxam, Gilbert, 1977 ; Sanger, Air et al., 1977), celui d'un génome, qui implique le traitement d'une grande masse de données et l'identification de l'organisation des gènes, utilise l'informatique comme l'outil privilégié pour accélérer et automatiser le processus de calcul qui aboutit à décrire en termes de gènes ce qui est donné, en entrée, en termes de suite de nucléotides (un gène est une suite délimitée de nucléotides). Du point de vue de la science informatique, les domaines mis en avant sont la théorie des langages et l'algorithmique. En effet, d'une part, les suites de nucléotides sont perçues comme des mots appartenant au « langage génétique » défini sur l'alphabet {A., C., G., T.} dont il faut décider s'ils correspondent ou non à des gènes (Hopcroft, Ullman, 1979). Pour déterminer si une séquence est codante, on peut utiliser des outils informatiques de prédiction capables d'identifier un gène selon plusieurs critères comme la présence d'une phase ouverte de lecture, de signaux d'épissage et la composition en bases. Divers logiciels ont été développés pour reconnaître les gènes sur la base d'informations compilées depuis toujours manuellement et regroupées « informatiquement » en bases de données, sources de vérifications et comparaisons. Loin d'être infaillibles, ces logiciels sont fondés sur l'utilisation de méthodes informatiques et mathématiques diverses comme les réseaux de neurones (McCulloch, Pitts, 1943 ; Bishop, 1995), l'analyse discriminante (Saporta, 2006) ou encore les méthodes de Monte-Carlo (Krauth, 1996), utilisant des chaînes de Markov. Parmi les plus connus, on trouve : GENSCAN (Burge, Karlin, 1997), un programme général de prédiction de séquences codantes à partir de séquences d'A.D.N. génomique ; FASTA (Lipman, Pearson, 1985 ; Pearson, Lipman, 1988), servant à trouver des séquences dans des bases de données et à identifier des structures périodiques basées sur des similarités de séquences locales ; et BLAST (Altschul, Gish et al., 1990), assez proche de FASTA au niveau de son objectif, existant en de nombreuses versions selon les besoins, qui permet de comparer des séquences données à des séquences connues.

Ces trois logiciels ne sont bien sûr pas les seuls pertinents dans le contexte de la bio-informatique des séquences. Certains ont en effet une autre utilité. À titre d'exemples, NNSPLICE a été développé sur la base des réseaux de neurones dans le cadre du Berkeley Drosophila Genome Project (http://www.fruitfly.org/) et permet de rechercher des sites d'épissage dans les séquences génomiques ; EST2GENOME (Mott, 1997), quant à lui, sert à la recherche d'exons et d'introns d'une séquence génomique par alignement avec des séquences d'A.D.N. complémentaire d'A.R.N.m. en tenant compte des consensus stricts des sites d'épissage.

Quelles que soient les méthodes informatiques et mathématiques au cœur de ces divers logiciels, la quantité croissante des données à traiter amène les acteurs de cette forme de bio-informatique à les développer toujours plus. En particulier, afin d'améliorer leurs performances en termes de fiabilité et de rapidité, une part importante des recherches actuelles s'organise autour de l'algorithmique des séquences. Ce domaine de l'informatique constitue l'un des développements, à l'instar de la combinatoire des mots par exemple, de la théorie des langages et vise à traiter algorithmiquement du texte au travers de problématiques de localisation de motifs, d'indexation, ou encore d'alignement de séquences (Crochemore, Hancart et al., 2001 ; Durbin, Eddy et al., 1998).

Du point de vue de la biologie, cette bio-informatique des séquences s'est avérée d'une importance capitale dans le cadre du séquençage des génomes et reste largement utile à l'heure actuelle en ce qui concerne l'aide à la comparaison de séquences. Sans l'informatique moderne et certaines méthodes qui ont été développées de manière *ad hoc* pour être appliquées à ces problèmes émanant de la biologie moléculaire,

le problème du séquençage de génomes à A.D.N. n'aurait certainement jamais trouvé de réponse, en raison de la quantité de données à traiter bien trop importante pour l'être humain démuni d'assistance automatisée. Plus précisément, nous pouvons aujourd'hui raisonnablement dire que le séquençage a été rendu possible par les développements purement informatiques autour du parallélisme (c'est-à-dire la possibilité d'avoir des architectures matérielles et logicielles capables de réaliser des opérations de manière simultanée afin de réduire les temps de calcul) (Jordan, Alaghband, 2002) et par les méthodes de *shotgun,* connues et utilisées dès la fin des années 1970 (Staden, 1979) et largement développées par Gene Myers et son groupe (Weber, Myers, 1997), articulées autour de deux stratégies : celle du séquençage aléatoire global (connue sous le nom de *whole genome shotgun*), pour des génomes de petites tailles (les progrès en informatique permettent aujourd'hui de l'utiliser sur des génomes de plus en plus importants), et celle dite "clone par clone" (connue sous le nom de *hierarchical shotgun*), plus adaptée aux génomes de (très) grande taille (Waterston, Lander et al., 2002). Ces techniques sont, comme leurs noms l'indiquent, toutes deux fondées sur la méthode *shotgun*. Cette méthode consiste à découper un génome en un grand nombre de fragments, de séquencer les extrémités d'une partie de ces fragments puis de les ré-assembler en fonction de leurs chevauchements afin de reproduire une séquence complète, ce qui soulève des difficultés spécifiques vis à vis du nombre de fragments permettant de recouvrir le génome dans son intégralité et de leur ré-assemblage. Ainsi, bien que ces deux méthodes aient largement été utilisées, la dernière a notamment été celle retenue par le Consortium international pour le séquençage humain, le *Human Genome Project* (ou *H.G.P.*). Par ailleurs, soulignons que de nouvelles techniques plus rapides et précises tout en étant moins coûteuses, dites de "haut débit", ont fait leur apparition depuis une dizaine d'années.

Au delà du séquençage, cette bio-informatique des séquences est au cœur des méthodes d'aide à la comparaison de séquence. Elle vise en effet aussi à développer des méthodes toujours plus fines et rapides. Aujourd'hui, les *S.N.P.* (*Single Nucleotide Polymorphism*) permettent de repérer les régions du génome comportant des variants génétiques (d'une seule paire de bases) chez des individus d'une même espèce. Fréquentes, ces variations se retrouvent dans les exons, les introns ou dans des régions inter-géniques et peuvent rendre compte de pathologies, comme la schizophrénie ou le diabète de type 2, grâce à la comparaison de la fréquence des allèles de ces *S.N.P*. chez des malades et des témoins réalisée dans des études d'association pangénomique (bien connues sous le nom de *G.W.A.S.*, pour *Genome Wide Association Studies*) (Bush, Moore, 2012). Ici, ces comparaisons mènent essentiellement à dégager des résultats statistiques, souvent fondés sur des corrélations, les tests d'hypothèse et le test du χ^2. Bien sûr, étant donné la très grande masse de données de comparaison à traiter, l'informatique revêt alors sa dimension technologique pour fournir un moyen de calcul automatisé permettant de réaliser l'ensemble des tests statistiques de manière efficace et fiable. Ces *G.W.A.S.*, combinées à la machinerie informatique sous-jacente, forment un véritable objet de découverte en permettant d'identifier des associations entre gènes ou mutations et pathologies, sans aucune hypothèse physiopathologique *a priori*. Par exemple, des travaux ont montré que les *G.W.A.S.* ont permis d'identifier plusieurs régions du génome humain portant des variants génétiques associés au diabète de type 2 (Billings, Florez, 2010), à l'obésité (Wang, Li et al., 2011) ou à l'infarctus du myocarde (Myocardial Infarction Genetics Consortium, 2009). Auparavant pourtant, aucun des gènes ou régions du génome mis en exergue par ces *G.W.A.S.*, n'était *a priori* considéré comme pouvant contribuer à ces pathologies. Plus étonnant encore, certaines régions identifiées ne semblent même pas contenir de gène ! Enfin, plus récemment, des développements informatiques autour de nouveaux logiciels tendent à rendre plus précises encore les analyses (Delaneau, Marchini et al., 2011), ouvrant ainsi des perspectives thérapeutiques encore plus ciblées (Limou, Zagury, 2013).

2.2 Bio-informatique des structures

Une part significative de ce qui entre dans la définition conventionnelle de la bio-informatique a trait à l'analyse, la compréhension et la prédiction des structures tridimensionnelles des (macro-)molécules

biologiques, comme par exemple l'A.D.N., l'A.R.N., les protéines ou encore les morphogènes ou hormones (Bourne, Weissig, 2003). Le problème général posé par la « bio-informatique structurale » est celui qui vise, à partir des séquences (cf. plus haut), à prédire et à analyser les structures macro-moléculaires qui en dépendent, afin de développer les intuitions quant à leur fonctionnalités spécifiques. Les recherches dans ce cadre visent donc à accroître la connaissance acquise sur les architectures spatiales des molécules et ont entre autres pour objectif de rendre possibles des comparaisons et des classifications entre molécules d'un même type (comme la comparaison de motifs locaux), d'établir les lois qui régissent certaines de leurs propriétés (par exemple celles qui sont à l'œuvre dans le phénomène de pliage moléculaire), des prédictions quant aux liens existant entre leurs composants atomiques. De telles recherches nécessitent de combiner les résultats issus des études en biologie moléculaire à ceux provenant des développements de l'informatique. Ainsi, sur la base des informations émanant notamment de la bio-informatique des séquences, de nombreuses questions sont actuellement posées sur les possibilités de prédire la structure des protéines à partir des informations connues sur les séquences d'A.D.N. et/ou d'A.R.N. qui les produisent. Les méthodes informatiques directement concernées ici sont principalement liées à la géométrie « computationnelle », en particulier discrète, et à l'algorithmique afin de développer les protocoles efficaces pour l'analyse des données cristallographiques et issues de spectroscopie de résonance magnétique nucléaire, afin de créer des modèles moléculaires valides. À titre d'exemple, de nombreuses études ont été menées sur les mesures de distances au sein des structures tri-dimensionnelles, grâce au développement parallèle des méthodes de géométrie des distances (Moré, Wu, 1999 ; Liberti, Lavor et al., 2008) et d'optimisation (Cutello, Narzisi, 2006).

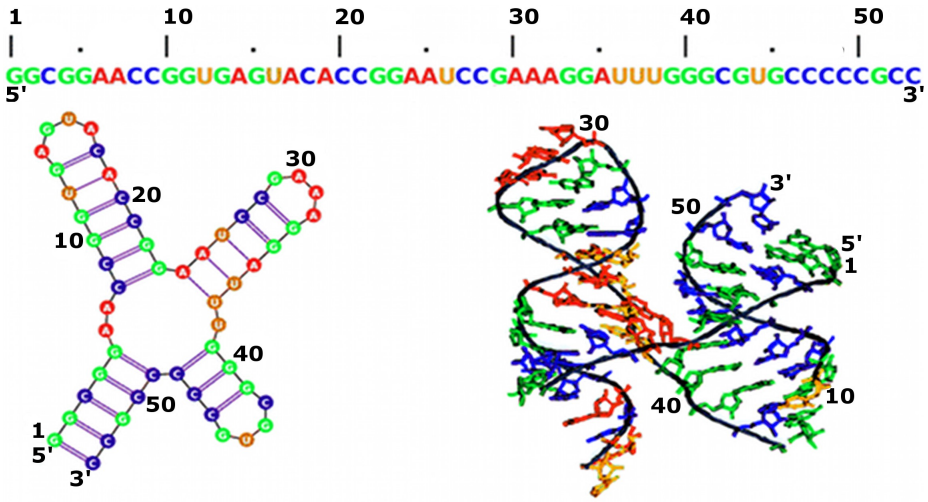

Fig. 1 : Structures primaire (en haut), secondaire (en bas à gauche) et tertiaire (en bas à droite) de l'A.R.N. référencé 1KH6 dans *Protein Data Bank* (image issue de (Kim, Fuhr et al., 2013)).

Afin de mettre en évidence l'intérêt de l'informatique dans ce domaine de recherche, revenons sur les A.R.N. et rappelons brièvement les concepts de structure de cette molécule. On parle de structure primaire pour désigner une séquence de nucléotides et de structure secondaire dès lors que sont identifiées les régions appariées. Enfin, les A.R.N. simple brin, lorsqu'ils se replient, le font le plus souvent sur eux-même et forment alors une structure intramoléculaire, dite tertiaire, qui peut s'avérer très stable et compacte (cf. Figure 1). Selon la structure tertiaire (tri-dimensionnelle) que les A.R.N. adoptent après leur repliement, ils peuvent se retrouver à l'origine de fonctions biologiques complexes et être ciblés par des molécules spécifiques, comme des antibiotiques ou des antiviraux. Les études portant sur ces structures revêtent par conséquent un intérêt particulier. Dans ce domaine, la bio-informatique structurale permet, par la visualisation et la manipulation des séquences issues des bases de données, la prédiction des caractéristiques de repliement des structures primaires menant à la compréhension de leurs structures secondaires et tertiaires. Or, la conformation fonctionnelle d'un A.R.N. dépend de son repliement en structure secondaire et également des interactions entre les régions appariées alors créées. Il découle alors que la fonction des

A.R.N. est analysable au travers de leur géométrie. De plus, il s'avère que certaines contraintes stériques, très souvent, autorisent de limiter l'ensemble des structures tertiaires possibles à des structures secondaires, naturellement plus simples car définies sur la base d'un moins grand nombre de paramètres. Dès lors que cela est rendu possible, l'ensemble des conformations d'un A.R.N. est représentable par un objet combinatoire « simple », que l'on peut assimiler par exemple à un arbre (cf. {Annexe 2} ; la hiérarchie des répertoires/dossiers et des fichiers dans un ordinateur est un exemple d'arbre) (Cormen, Leiserson et al., 1994) ou à un mot de Dyck (c'est-à-dire un mot bien parenthésé) (Carton, 2008). L'avantage d'une nouvelle représentation de ce type est qu'il est alors possible, pour mener à bien l'étude fonctionnelle d'un A.R.N., d'utiliser l'ensemble des méthodes bien connues issues de la combinatoire analytique.

Si l'on voulait résumer grossièrement, l'on pourrait dire que la bio-informatique des structures consiste, au moyen de la science et de la technologie informatique, à reconstruire les structures, effectuer des prédictions structurales et analyser les propriétés dynamiques de macro-molécules, en se reposant souvent sur les descriptions existantes des forces moléculaires qui agissent sur les atomes pour simuler des mouvements et trouver des conformations d'énergie minimale. À cet égard, il est connu que le repliement des A.R.N. est affecté par la température du milieu environnant. Des méthodes, issues de la bio-informatique structurale, ont ainsi été développées pour prédire l'influence des changements de température sur la régulation de l'expression génétique (Brierley, Gilbert et al., 2008). Cette bio-informatique est aussi centrale dans le cadre de l'étude et de l'identification des séquences cibles d'A.R.N. régulateurs (par exemple les micro-A.R.N et les A.R.N. interférents) et des complexes protéiques.

2.3 Bio-informatique des réseaux

Les éléments qui jouent un rôle moteur dans la spécification, l'organisation et la dynamique des individus sont nombreux (gènes, A.R.N., protéines, hormones, neurones, etc.). Au delà de ce paramètre numérique qui est l'une des sources de la complexité du vivant, on ne peut aujourd'hui penser comprendre le vivant sans comprendre les interactions qui s'exercent entre ces éléments. Très souvent, en biologie, on appelle ces interactions des régulations, que nous représentons sous la forme de réseaux.

En termes plus formels, ces réseaux sont généralement représentés par des graphes (Berge, 1958) dont les sommets représentent les éléments biologiques en question (par exemple des gènes) et les arêtes (orientées ou non, c'est-à-dire indiquant ou non le rôle de l'un par rapport à l'autre) correspondent aux actions effectives des éléments les uns sur les autres (cf. {Annexe 1}). Plus précisément, dans le cas des réseaux de régulation génétique, les arêtes sont généralement orientées, on les appelle alors aussi des arcs, et signées, le signe représentant le caractère inducteur ou inhibiteur de la régulation du gène source sur le gène cible. À titre d'illustration, la Figure 2 représente la formalisation sous forme de graphe d'interaction d'un des réseaux de régulation génétique admis du contrôle immunitaire chez le phage[3] lambda mettant en avant les régulations entre les gènes cI, Cro, N et cII (Thieffry, Thomas, 1995). De manière générale, sur ces réseaux ressortent de la littérature des études qui peuvent être classées selon deux axes. Le premier se place en amont et est essentiellement théorique. Il vise à utiliser des techniques issues des mathématiques discrètes et de l'informatique fondamentale afin de découvrir et comprendre les liens qui existent entre l'architecture des réseaux et leur propriétés dynamiques sous-jacentes. Dans ce cas, les réseaux sont eux-mêmes essentiellement théoriques, dans le sens où ils ne proviennent pas de données biologiques, l'objectif étant de mettre en emphase les conditions indispensables aux réseaux biologiques pour que ces derniers possèdent les attributs de complexité et dynamiques généralement retrouvés dans le vivant. Le second, en aval, et plus appliqué et se rapproche de la biologie puisqu'il vise à analyser les données biologiques issues des

3 Un phage est un virus porté par une bactérie. Certains phages ont la particularité de pouvoir vivre dans des bactéries sans influencer leur comportement (en sommeil sous la forme de prophage), et peuvent se réveiller à certains moments, ce qui peut entraîner la mort de la bactérie qui les contient.

expérimentations et accessibles depuis la littérature afin d'en tirer des réseaux dits réels, qui sont dès lors étudiés au moyen des mêmes méthodes que celles servant le premier axe, provenant principalement de la théorie des systèmes dynamiques et de la complexité.

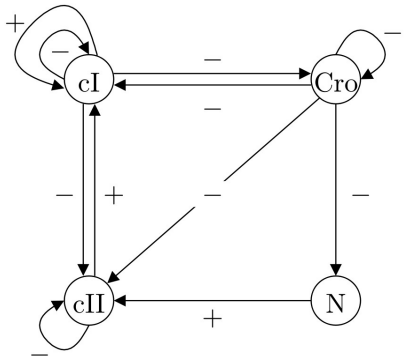

Fig. 2 : Graphe d'interaction représentant un réseau de régulation génétique modélisant le contrôle immunitaire dans le phage lambda.

L'une des premières régulations génétiques fonctionnelles à avoir été décrite est celle de l'opéron lactose de la bactérie *Escherichia Coli* par Jacob et Monod, qui permet notamment de comprendre les échanges de gènes entre bactéries. Dans les années 1950, Lwoff étudie à l'Institut Pasteur la croissance des bactéries porteuses de phages (Lwoff, Siminovitch et al., 1950 ; Lwoff, 1959). Il découvre que la destruction des bactéries ne dépend pas uniquement de la nature des phages qu'elles portent mais également d'un agent extérieur capable de moduler l'activité des phages (en l'occurrence un rayonnement ultraviolet). C'est alors que Jacob, accueilli par Lwoff, approfondit l'étude de ce phénomène dans sa thèse de doctorat (Jacob, 1954). En parallèle, Monod étudie la croissance bactérienne. Pour vivre, les bactéries produisent des enzymes capables de digérer des sucres dont elles se servent comme source d'énergie. En particulier, Monod s'aperçoit que les enzymes nécessaires à la consommation de lactose ne sont produites qu'en présence de ce sucre dans le milieu de croissance. Il en conclut que la synthèse de ces enzymes est induite par le lactose et peut varier au cours du temps (Monod, Wollman, 1947 ; Monod, 1958). En croisant leurs études respectives, Jacob et Monod établissent un modèle original, représentable formellement sous forme de réseau, permettant d'expliquer le « système lactose » et le réveil des phages (Jacob, Perrin et al., 1960 ; Jacob, Monod, 1961). Bien sûr, les recherches de ces trois hommes sur le contrôle génétique des synthèses virales et enzymatiques sont récompensées en 1965 par le prix Nobel de médecine mais, plus important, elles représentent la première pierre en biologie moléculaire de ce qui pourrait être appelé aujourd'hui la science de la régulation[4], cette dernière étant certainement le cœur de la biologie.

La découverte et la formalisation de tels réseaux de contrôle permet de représenter formellement de manière abstraite les régulations. Ces réseaux sont alors des objets mathématiques complexes, à savoir des graphes d'interaction, qui permettent d'approximer la réalité biologique en se libérant d'un certain nombre de paramètres (dont la prise en compte entraînerait une complexité qui rendrait toute analyse irréalisable) tout en en conservant l'essence. Les graphes d'interaction résultant modélisent alors l'aspect statique des régulations qui peut être étudié pour lui-même et qui possède généralement un caractère dynamique qui, lui-même, peut également être analysé par des méthodes largement développées depuis longtemps aussi bien en informatique qu'en mathématiques. L'utilité pour la biologie vient de cette modélisation, à l'origine de la simplification analytique des lois du vivant permettant d'acquérir les conditions nécessaires (mais généralement non suffisantes) pour en comprendre le fonctionnement. Un exemple illustrant cette modélisation est celui des voies d'activation moléculaires, ou voies de signalisation, qui ont fait l'objet de représentations de plus en plus complexes et qui se trouvent aujourd'hui « miniaturisées », analysées une par

4 La science de la régulation a en fait été développée avant dans d'autre discipline telles que l'informatique, malgré un champ sémantique différent où le terme régulation est souvent remplacé par celui d'interaction.

une et ré-intégrées dans un contexte « pathologie-dépendant » (comme le cancer ou l'inflammation). Ce type d'études mène également à de nombreux développements dans le domaine des bio-technologies. En effet, certaines entreprises de matériel pour la recherche en biologie proposent des kits utilisables pour identifier, dans des cellules en culture, la dose optimale pour un traitement, la réponse des cellules cancéreuses à l'inhibition d'une voie de transduction du signal, ou encore pour mesurer la signalisation en réponse à une cytokine.

Afin d'illustrer concrètement ce domaine de la bio-informatique des réseaux, prenons en considération le réseau de l'adhésome, à savoir la carte des gènes acteurs du processus d'adhérence cellulaire, tel qu'il est évoqué dans (Zaidel-Bar, Itzkovitz et al., 2007). L'adhérence cellulaire dépend à la fois des composants de la matrice extra-cellulaire et de la cellule elle-même, et en particulier de ses récepteurs membranaires. Notamment, la cellule met en jeu des réponses adaptées aux perturbations mécaniques engendrées par la nature physique de son environnement, en « détectant » par exemple la rigidité relative de la surface sur laquelle la cellule se trouve (Lo, Wang et al., 2000 ; Zaidel-Bar, Kam et al., 2005). Afin de mieux comprendre les mécanismes qui sous-tendent ces réponses cellulaires, des groupes de recherches se sont associés sous la forme d'un projet (http://www.adhesome.org) avec le support du *N.I.H.*, pour construire un réseau *in silico* modélisant certaines des régulations de l'adhésome. Cette construction est bien sûr passée par la découverte expérimentale des protéines impliquées et des voies métaboliques activées (Brown, Turner, 2004 ; Zaidel-Bar, Kam et al., 2005 ; Legate, Montañez et al. 2006) et la collecte, depuis la littérature, de l'information sur les composants pouvant participer à l'adhésome (Zamir, Geiger, 2001 ; Lo, 2006). Il en ressort une étude assez complète mettant en exergue les composants connus de la matrice extra-cellulaire, menant à la description d'un réseau de régulation bio-chimique. En se fondant sur les travaux passés d'Alon (Alon, 2003, 2007), elles-mêmes fondées sur la recherche de sous-graphes dans des graphes et des méthodes de classification statistiques, le projet *Adhesome* insiste sur la présence dans ce réseau de petits motifs récurrents bien connus pour être présents en grand nombre dans la plupart des réseaux de régulation, tels que les motifs antérogrades, les motifs de rétroaction, ou encore les *bifans* (cf. Figure 3).

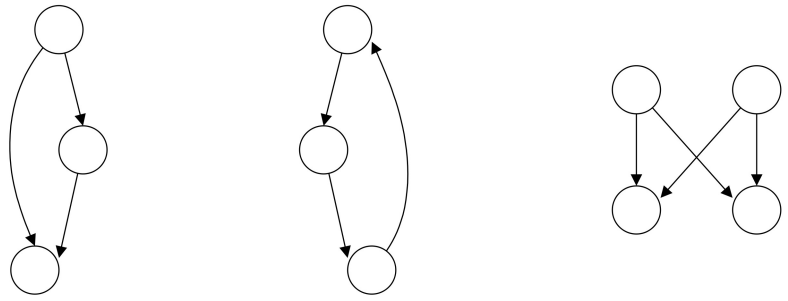

Fig. 3 : Trois motifs récurrents dans les réseaux de régulation biologiques ; *(a)* un motif de régulation antérograde, *(b)* un motif de rétroaction, ou cycle, *(c)* un *bifan*.

Ce couplage entre le travail expérimental et le travail bibliographique menant à la construction d'un réseau est l'une des clés de la bio-informatique des réseaux dans sa dimension applicative. Ici, il permet de mieux appréhender comment la machinerie cellulaire « ressent » les signaux environnementaux et leur répond. Cependant, il apparaît aujourd'hui évident du point de vue de la modélisation et de l'informatique que ce type d'étude est loin d'être suffisant pour espérer expliquer les mécanismes d'adhérence cellulaire. Cela est également vrai d'un point de vue plus général pour plusieurs raisons. Parmi elles, évoquons tout d'abord le fait que la construction des réseaux ne peut se faire sans déconstruction (généralement partielle), suivie d'une reconstruction, et ainsi de suite. La question, une fois la première construction terminée est avant tout de s'enquérir de la validité du modèle, dans une certaine mesure tout du moins. Or, dès lors que l'on pose la question de la validité d'un modèle, l'on écarte inévitablement son regard de la partie émergée

de l'iceberg, pourtant indispensable, qu'est la vision des aspects statiques de la régulation (par exemple, l'étude évoquée plus haut sur l'adhésome), pour l'orienter vers sa partie immergée que représentent ses aspects dynamiques, dont la complexité entraîne celle du vivant. En effet, seules des connaissances approfondies sur la dynamique d'un réseau, associées à celles issues d'expérimentations et de la littérature, peuvent permettre d'en distinguer les parties valides et cela passe en général par de nombreux allers-retours entre biologistes et informaticiens.

2.4 Traitement de l'information

Nous ne pouvons pas raisonnablement parler de bio-informatique sans évoquer le traitement de l'information. Au vu de la masse de données accumulée en biologie depuis des décennies, notamment depuis l'utilisation des techniques informatiques dans les domaines présentés plus haut (séquençage, puces à A.D.N....), l'un des défis posés à la bio-informatique a été (et est encore aujourd'hui) la gestion de ces données, au sens large. Ce défi est d'autant plus important que l'accroissement des données se poursuit exponentiellement. Heureusement, indépendamment de la biologie et de la bio-informatique, la science informatique s'intéresse depuis longtemps aux différentes questions liées au traitement de l'information. En particulier, de nombreuses recherches ont vu le jour autour de questions ayant trait au stockage et à l'organisation (aspect syntaxique du traitement) des données ainsi qu'autour de celles relatives à leur réorganisation et leur étude à proprement parler pour en tirer du sens (aspect sémantique du traitement). Ces familles de questions appartiennent respectivement aux domaines des *bases de données* et du *data mining*. Ici, nous n'allons pas insister sur les bases de données car les méthodes mises en œuvre sont purement informatiques et ne dépendent aucunement de la nature des données à traiter. À titre d'information cependant, les bases de données couramment utilisées à ce jour sont des bases de données relationnelles dont les fondements ont été introduits par Codd (Codd, 1970). Le *data mining*, quant à lui, prend un sens particulier en bio-informatique étant donné que l'accumulation de données amène à remplir toujours plus les bases de données qui deviennent *de facto* telles que le ratio connaissance/information est en constante diminution relative. En se plaçant à la croisée de nombreux domaines de l'informatique, comme l'algorithmique, l'apprentissage automatique et statistique, la représentation (visualisation) des connaissances..., il représente le processus qui vise à extraire de la connaissance, ou plus précisément des motifs intéressants (non triviaux et généralement implicites), à partir de grands volumes de données « brutes ». Le processus de *data mining* peut être séparé en deux phases : la première concerne la préparation des données et vise à collecter, nettoyer, intégrer, transformer et filtrer les données pertinentes pour le problème posé, la seconde consiste quant à elle à explorer les données ainsi préparées en vue de leur analyse, qui s'oriente *a posteriori* vers la prédiction de modèles spécifiques de systèmes biologiques réels (Chen, Lonardi, 2009 ; Hall, Frank et al., 2009), qui peuvent être des modèles de structures d'A.R.N., de réseaux...

Bien sûr, comme nous l'avons dit, le développement des méthodes de traitement des données est sans aucun doute essentiel à celui de la bio-informatique moderne, que cette dernière soit vue dans n'importe laquelle des formes qu'elle peut revêtir et qui ont été développées plus haut. Toutefois, le traitement des données est un thème de recherche à part entière, qui ne dépend pas dans ses aspects fondamentaux de la nature des données elles-mêmes mais de leur forme. C'est pourquoi nous n'allons pas le détailler plus avant dans ce chapitre.

2.5 Une application traditionnelle : la phylogénie

La phylogénie est un domaine de recherche qui s'inspire des précédents et qui vise à analyser les relations de parenté entre des organismes. Les organismes à l'étude peuvent soit appartenir à des espèces

différentes, soit à une même espèce. Dans tous les cas, l'objectif général est de parvenir à mieux comprendre les propriétés d'évolution des organismes vivants au cours du temps et de retracer « l'histoire » des espèces. Les relations de parenté entre organismes sont classiquement représentées sous forme d'arbre (cf. Figure 4). Darwin, dans son ouvrage consacré à L'origine des espèces, paru en 1859, évoquait le fait que "*les affinités de tous les êtres de la même classe ont parfois été représentées sous la forme d'un grand arbre. Je crois que cette comparaison est très juste. Les rameaux verts et bourgeonnants peuvent représenter les espèces existantes; les branches produites les années précédentes peuvent représenter la longue succession des espèces éteintes."* Cette notion d'arbre donne un rôle particulier à l'informatique en raison de la forte utilisation des arbres pour représenter les données au sein des ordinateurs (cf. {Annexe 2}). On retrouve ainsi dans la littérature de nombreux ouvrages de mathématiques et d'informatique sur les arbres phylogénétiques (Semple, Steel, 2003 ; Gascuel, 2005 ; Gascuel, Steel, 2007).

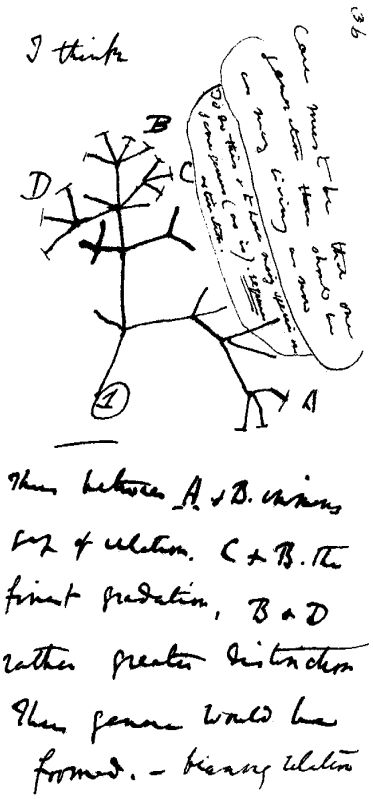

Fig. 4 : Premier schéma de Darwin représentant un arbre phylogénétique (Darwin, 1887). Interprétation du texte : « I think case must be that one generation should have as many living as now. To do this and to have as many species in same genus (as is) requires extinction . Thus between A + B the immense gap of relation. C + B the finest gradation. B+D rather greater distinction. Thus genera would be formed. Bearing relation » (début de la page suivante) « to ancient types with several extinct forms ».

Outre leur pendant informatique principalement fondé sur la conception et l'implantation d'algorithmes toujours plus efficaces pour la construction des arbres phylogénétiques sur lequel nous reviendrons plus bas, les recherches en phylogénie ont un champ d'application vaste qui couvre l'ensemble des aspects liés à la biodiversité. En particulier, elles visent à mieux appréhender l'ensemble des facteurs influençant son évolution. Si l'on se focalise par exemple sur l'agriculture, l'on s'intéresse notamment à l'évolution des spectres d'hôtes de parasites ravageurs. Dans un contexte de santé publique, les questions d'intérêt portent plus spécifiquement sur l'évolution des pathogènes en relation avec celles de leurs hôtes, de leur vecteurs ou encore de leurs réservoirs. Nous n'avons jusque là évoqué que le niveau macroscopique des recherches en phylogénie (à savoir celui des organismes et de leurs interactions). Toutefois, il convient de préciser que le niveau microscopique a lui aussi été au centre de nombreuses études, en particulier ces

dernières années. Ainsi, des développements remarquables ont été faits en phylogénie moléculaire, qui vise à analyser les liens de parenté entre des séquences de nucléotides ou d'acides aminés, autrement dit l'évolution des génomes (Lopez, Casane et al., 2002). Ils ont notamment permis d'établir une relation entre les mutations neutres qui n'affectent pas les capacités de reproduction des individus et les différences des caractéristiques génétiques des espèces admettant un ancêtre commun (Page, Holmes, 1998). De telles avancées permettent de construire l'arbre phylogénétique des distances entre séquences, qui renseigne sur les lignées des espèces, comme Darwin l'avait lui-même pensé (cf. plus haut).

Revenons à présent sur les apports de l'informatique dans le contexte des recherches en phylogénie. Ils peuvent être de deux sortes, énumérées ci-dessous :

- le premier type cherche à trouver les méthodes de traitement de données d'évolution les plus efficaces, voire optimales, permettant d'aboutir à la constitution d'arbres phylogéniques. Dans ce cadre, il existe principalement trois méthodes (Felsenstein, 1996) menant à la construction de ces arbres, qui, font essentiellement appel à des techniques algorithmiques afin d'optimiser la construction en termes de complexité en temps et en espace. La première est fondée sur le principe des *distances*. Elle vise à prendre en compte dans la construction les distances entre les feuilles de l'arbre, à savoir les éléments de ce dernier qui représentent des organismes contemporains. Ici, deux algorithmes peuvent être mis en avant : le "*unweighted pair group method with arithmetic mean*" (Sokal, Michener, 1958 ; Murtagh, 1984) qui a progressivement été remplacé par le "*neighbour joining*" (Saitou, Nei, 1987) pour des raisons d'efficacité. La deuxième méthode est connue sous le nom de méthode par *parcimonie*. L'idée générale est de retenir parmi les arbres satisfaisant les hypothèses évolutionnaires, celui qui optimise le nombre de substitutions (Fitch, 1971 ; Felsenstein, 1978). Une telle méthode peut s'avérer très coûteuse selon le nombre d'arbres comparés. La dernière méthode est celle dite de *vraisemblance*. Fondée sur une approche probabiliste, elle vise à construire des arbres phylogénétiques en fonction des taux de substitutions des éléments de base sur lesquels les arbres sont déduits, en calculant la position, la longueur et les relations entre les branches des arbres (Lewis, 1998 ; Guindon, Gascuel, 2003 ; Guindon, Dufayard et al., 2010).
- l'autre est sans aucun doute plus prospectif et vise à analyser plus avant les "messages" délivrés (bien que très souvent cachés) par ces arbres, en combinant les informations qu'ils contiennent à celles issues d'autres études de nature biologique, mathématique ou encore informatique. Il s'agit alors de modéliser ces messages, afin d'en tirer une synthèse en vue d'un traitement ultérieur permettant d'en apprendre plus sur le domaine.

En guise de conclusion sur cette application récurrente de la bio-informatique, nous évoquons maintenant l'une des ouvertures qui nous semble parmi les plus pertinentes et innovantes dans le cadre de la phylogénie et qui pourrait accroître considérablement la connaissance de l'évolution des espèces : la phylogénie des réseaux. Les études menées en phylogénie jusqu'à maintenant se sont essentiellement focalisées sur les gènes en tant qu'objets d'étude à proprement parler sans se concentrer sur les régulations entre ces derniers. Or, il est indéniable aujourd'hui que connaître le rôle des gènes en tant qu'éléments propres d'un réseau de régulation n'est pas suffisant pour comprendre les relations entre les organismes. Il est par conséquent probable, mais surtout essentiel, que les recherches futures s'orientent vers la phylogénie des réseaux de régulation. De telles études permettront à n'en pas douter de mettre en avant de nouveaux concepts à la frontière de la biologie, des mathématiques et de l'informatique. C'est notamment l'un des principaux fils directeurs des recherches menées par Alon (cf. partie 2.3) et son groupe autour des motifs récurrents dans les réseaux de régulation. Il va de soi que comprendre statiquement comment les régulations ont évoluées au cours de l'évolution est un élément phare dans l'analyse de l'histoire des espèces. Toutefois, comme nous l'avons déjà expliqué, nous pensons que cela est loin d'être suffisant car la complexité du vivant ne réside pas seulement au niveau statique des régulations. Loin s'en faut. Intrinsèquement, cette complexité émane de la dynamique des régulations, causé en partie par l'architecture des régulations mais aussi par de nombreux autres paramètres qu'il est par ailleurs impossible d'énumérer ici de façon exhaustive. Pour comprendre les caractéristiques fondamentales de cette complexité, il est nécessaire de s'abstraire d'un sous-ensemble conséquent de ces paramètres et de se rapprocher de l'informatique et des mathématiques. Sur

ce sujet justement, des études théoriques récentes (Noual, 2012 ; Sené, 2012) ont donné de premières pistes à suivre, qui mettent en avant le rôle singulier de certains motifs de régulation génétique extrêmement simples du point de vue architecturel et qui ont la capacité de reproduire des comportements extrêmement complexes (ce qui se traduit notamment par un nombre de comportements exponentiel). Ces études préliminaires posent naturellement la question de l'évolution des réseaux. Elles mettent en avant la possibilité que la nature serait partie d'organismes simples architecturellement au niveau des régulations génétiques et très complexes au niveau de leur expressivité en termes de fonctions biologiques pour les transformer, au fil de l'évolution, en des organismes complexes architecturellement et néanmoins beaucoup plus simples fonctionnellement (les degrés de libertés comportementaux étant drastiquement réduits), en raison notamment des contraintes et des spécialisations successives subies. Sans aller plus loin, il convient de dire que ce type d'étude n'en est qu'à son balbutiement, qu'il n'est actuellement étayé par aucune donnée réelle tangible et qu'il ne s'agit là que de spéculation biologique néanmoins fondée mathématiquement.

3. Réseaux et modélisation

Dans cette partie, nous évoquons la modélisation des réseaux biologiques en nous focalisant sur deux aspects, le premier est théorique et le second appliqué. Le but est de poser dans un premier temps certaines des bases formelles des recherches actuelles sur les régulations biologiques et dont l'objectif est de caractériser les conditions nécessaires et/ou suffisantes pour que les réseaux possèdent les caractéristiques indispensables à la complexité intrinsèque des organismes vivants. Bien qu'elles ne soient pas abordées ici, les applications réelles sont larges et bien connues, aussi bien en biologie animale que végétale. Dans un deuxième temps, nous présentons des exemples de modélisation mettant en avant d'autres théories et méthodes informatiques afin d'étayer le discours et de montrer l'utilité de la modélisation dans le cadre de la compréhension de phénomènes biologiques.

3.1 Modélisation théorique : le cas des réseaux d'automates

Comme nous l'avons signalé dans l'introduction de ce chapitre, l'histoire montre que les liens entre la biologie et l'informatique sont anciens et remontent à la source même de l'informatique moderne. Sans vouloir aller vers une dichotomie de la bio-informatique contemporaine, il convient de comprendre que l'on peut l'appréhender de différentes manières. Ici, nous insistons sur le fait que la bio-informatique n'est pas uniquement une discipline qui vise à utiliser des méthodes informatiques sur des donnés biologiques. Elle peut en effet se situer à des niveaux d'abstraction supérieurs et se détacher des données biologiques réelles afin de découvrir « les grandes lois qui gouvernent le vivant ». Dans ce contexte, la toute première modélisation de réseaux biologiques remonte à 1943 avec les travaux sur les réseaux de neurones formels (McCulloch, Pitts, 1943). Ensuite, les recherches de Kauffman et Thomas à partir de la fin des années 1960 (Kauffman, 1969 ; Thomas1973) ont largement contribué à l'essor de cette thématique telle qu'on la connaît aujourd'hui en mettant notamment en avant l'importance des mathématiques et de l'informatique dans ce domaine. L'idée originale portée par Kauffman et Thomas (et d'autres, comme Demongeot et Goles) est de réussir à comprendre, sur la base d'informations essentiellement théoriques potentiellement éloignées de la biologie, quelles sont les propriétés que de tels réseaux doivent posséder pour être capables de reproduire la complexité du vivant. Ainsi, de nombreuses études sont intrinsèquement liées à des domaines fondamentaux de l'informatique comme la théorie des graphes et des systèmes dynamiques discrets, la complexité et la calculabilité, les réseaux de régulation pouvant naturellement être vus comme des machines abstraites permettant des calculs (la sélection d'un type cellulaire particulier par un réseau de régulation génétique étant dans ce cas le résultat du calcul effectué par le réseau lui-même). Afin de clarifier les propos précédents, nous allons dans cette partie présenter une approche, dite des réseaux d'automates, qui est au cœur de cette

partie de la bio-informatique et qui, nous pensons, offre tous les attributs utiles pour atteindre l'objectif évoqué.

Les réseaux d'automates sont des objets mathématiques discrets permettant de modéliser n'importe quel système d'entités en interaction, comme notamment les réseaux de régulation biologique. Dans ces réseaux, les entités, ou automates, admettent plusieurs états possibles. À titre d'exemple, en considérant que les gènes peuvent être exprimés ou non, on peut représenter des réseaux de régulation génétique au moyen de réseaux d'automates *booléens*, à savoir des réseaux dans lesquels tous les automates possèdent deux états possibles, 0 pour préciser la non-expression d'un gène et 1 pour son expression effective. Usuellement, on représente les réseaux d'automates par des graphes, appelés graphe d'interactions, qui sont tels que l'ensemble des sommets représente l'ensemble des automates à l'étude et que l'ensemble des arêtes représente l'ensemble des interactions entre ces automates. La seule connaissance du graphe d'interaction ne suffit toutefois pas à définir un réseau puisqu'un tel graphe ne donne d'informations que sur l'existence d'interactions entre les automates qui le composent et ne précise aucunement la nature de ces interactions. Pour remédier à cela, on définit généralement un réseau d'automates au moyen d'un ensemble de fonctions locales de transition, une pour chacun des automates. Chaque fonction détermine la manière dont l'état de l'automate auquel elle correspond évolue au cours d'un temps discret selon les états de ceux qui l'influencent. Plus formellement, si l'on travaille sur un réseau d'automates constitué de n automates (on dit alors qu'il est de taille n), celui-ci est défini par un ensemble de n fonctions locales de transition. Pour illustrer ces notions, la Figure 5 représente un réseau d'automates booléens de taille 3 et le graphe d'interaction qui lui est associé.

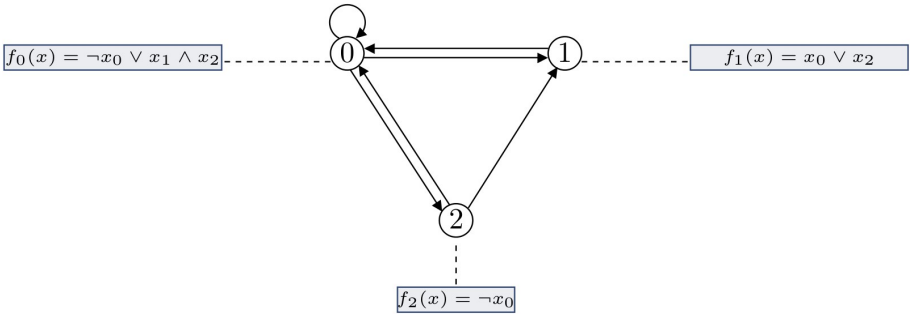

Fig. 5: Exemple d'un réseau d'automates booléens arbitraire de taille 3, présenté au travers l'ensemble des fonctions locales de transition qui le caractérisent et le graphe d'interaction qui leur est associé.

Ces réseaux, même réduits à l'abstraction booléenne, sont particulièrement adaptés à la modélisation des réseaux de régulation biologique. Tout d'abord, ils possèdent toutes les caractéristiques permettant de représenter ce qu'indiquent les biologistes au travers d'un langage propositionnel, comme « si tel et tel gènes sont tous deux exprimés, alors tel autre gène ne s'exprime pas » ou encore « si tel gène est exprimé ou que tel autre gène ne l'est pas, alors un troisième est exprimé ». Cette propriété d'expressivité du langage propositionnel par de tels réseaux a été démontrée par McCulloch et Pitts dans leur article original. Ensuite, ces réseaux possèdent les mêmes capacités de calcul que les machines de Turing universelles, ce qui signifie qu'ils permettent de calculer l'ensemble des fonctions calculables (sous réserve que l'on s'autorise un nombre infini d'automates). Pour illustrer ce point, prenons en considération les réseaux d'automates booléens. Malgré leur haut niveau d'abstraction qui mène inévitablement à mener des études de natures qualitatives plutôt que que quantitatives, ces derniers permettent de modéliser n'importe quel système d'entité en interaction. Par exemple, il est possible de modéliser un réseau de régulation génétique de telle manière que les sous-ensembles d'états possibles des gènes qui le composent dépend de la concentration des protéines qu'il produisent. Il suffit de représenter un gène par plusieurs automates. En effet, par une procédure d'encodage/décodage, k automates booléens permettent de représenter 2^k configurations distinctes, une configuration correspondant à l'instanciation d'un état, 0 ou 1, à chacun des automates. Par exemple, le

réseau de la Figure 5 admet 2^3 configurations possibles qui correspondent à tous les nombres de 0 à 7 encodés en base 2, à savoir 000, 001, …, 111.

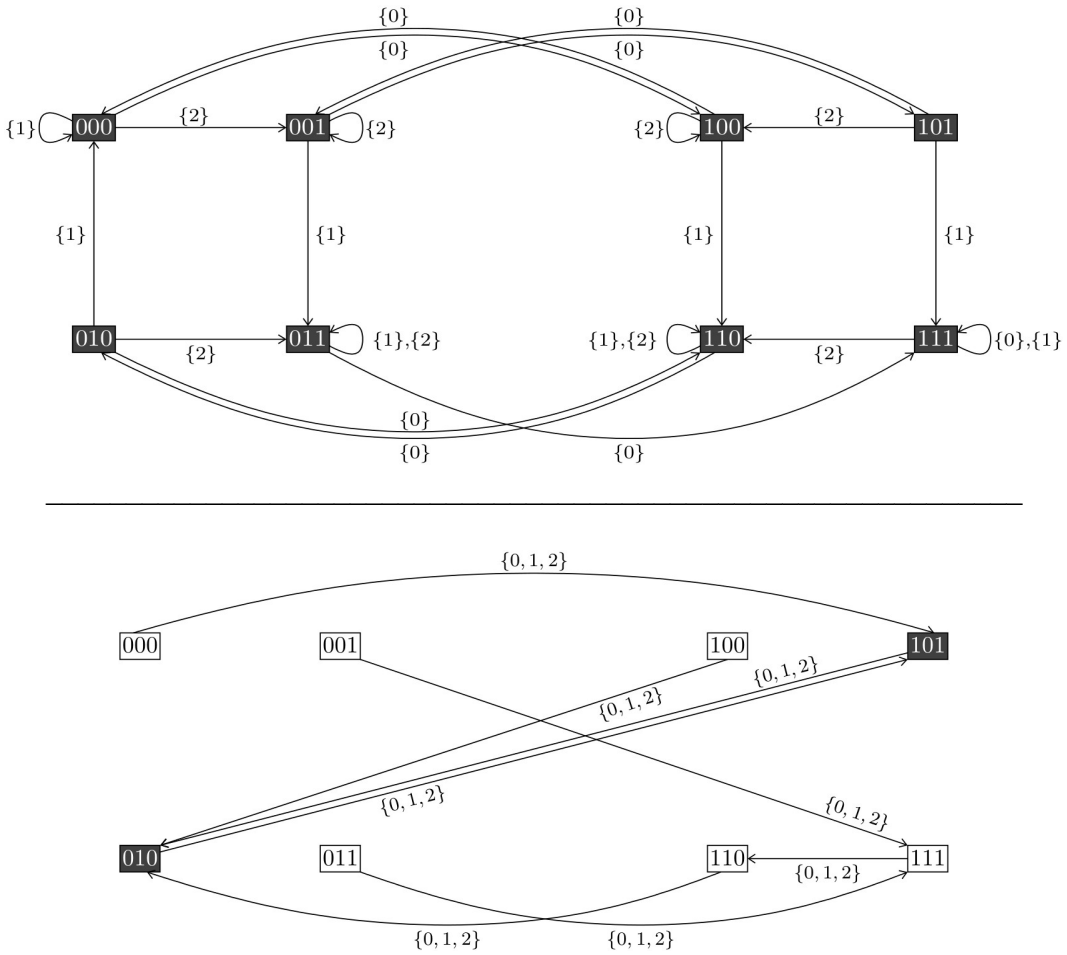

Fig. 6: Graphes de transition asynchrone (haut) et parallèle (bas) représentant deux systèmes dynamiques associés au réseau d'automates défini dans la Figure 5.

Le concept de configurations des réseaux d'automates est à rapprocher de celui d'état d'un réseau biologique. En effet, si l'on admet que le réseau de la Figure 5 représente une portion d'un réseau de régulation génétique réel, la configuration 101 représente un état du système biologique sous-jacent dans lequel seuls les gènes identifiés par 0 et 2 sont exprimés. Cela nous amène naturellement à évoquer les aspects dynamiques de ces réseaux. Nous avons précédemment introduit les fonctions locales de transition. Elles permettent de mettre à jour les états des automates en fonction des états de leurs voisins (le voisin d'un automate *i* étant un automate *j* tel qu'il existe une arête de *j* vers *i* dans le graphe d'interaction). Or, dès que l'on s'intéresse à la dynamique d'un réseau, les objets de référence ne sont plus les automates mais les configurations du système qui leur sont associées. Depuis une configuration, il existe plusieurs manières de mettre à jour les états des automates. On peut procéder de manière parallèle[5] (Kauffman, 1969), de manière asynchrone[6], comme l'a prôné Thomas (Thomas, 1991), ou encore selon n'importe quelle autre méthode, comme la méthode bloc-séquentielle[7] de Robert (Robert, 1995) ou encore la méthode générale[8] de Noual (Demongeot, Goles et al., 2010 ; Noual, 2012 ; Sené, 2012). Chacune de ces méthodes (ou modes) de mise à jour amène à représenter un système dynamique discret spécifique associé au réseau d'automates. Les

5 Les états de tous les automates du réseau sont mis à jour à chaque étape de temps.

6 Depuis chaque configuration, on peut mettre à jour l'état de chaque automate.

7 Les automates du réseau sont séparés en blocs disjoints. À l'intérieur d'un bloc, les automates sont mis à jour en parallèle, et les blocs eux-mêmes.

8 Depuis chaque configuration, on peut mettre à jour les états de tous les sous-ensembles d'automates.

systèmes dynamiques relatifs à un réseau d'automates sont généralement représentés par un *graphe de transition*, où l'ensemble des sommets représente celui des configurations du réseau et l'ensemble des arêtes orientées représente l'ensemble des transitions possibles entre les configurations. En d'autres termes, dans ce cadre, un graphe de transition représente l'évolution d'un réseau d'automates au cours du temps, chaque transition (x,y) modélisant une transformation, potentiellement non atomique, qui permet de passer de la configuration x à la configuration y. La Figure 6 illustre les graphes de transition asynchrone et parallèle du réseau défini dans la Figure 5.

En considérant les réseaux d'automates comme des modèles discrets qualitatifs des réseaux de régulation, leur comportement dynamique permet de représenter celui des réseaux de régulation. À ce sujet, les graphes de transition offrent des informations intéressantes. En particulier, lorsque les réseaux sont de taille finie, ce qui est le cas en biologie, leurs configurations sont nécessairement attirées vers des sous-ensembles de configurations particuliers qui se répètent indéfiniment. Au regard des graphes de transition, cela signifient qu'ils admettent de manière certaine des composantes fortement connexes terminales (une composante fortement connexe est un ensemble de sommets tels que, depuis chaque sommet la composant, il est possible d'accéder par un chemin composé d'une succession d'arêtes à tout autre sommet la composant également; on dit d'une composante fortement connexe qu'elle est terminale si aucune arête ne permet d'en sortir). Ces composantes forment ce qu'on appelle les *attracteurs* du réseau. Elles se résument soit à une unique configuration, auquel cas on appelle ce comportement asymptotique une *configuration stable*, soit à un ensemble de configurations dont le cardinal est strictement supérieur à 1, auquel cas on parle *d'oscillation stable*. Notons que les configurations stables restent stables quel que soit le mode de mise à jour choisi, ce qui n'est pas le cas des oscillations stables, comme le montre la Figure 6, dans laquelle les configurations appartenant à une oscillation stable sont représentées sur fond gris foncé.

L'intérêt des attracteurs des systèmes dynamiques (au sens large) en biologie a été clairement identifié par Delbrück dès la fin des années 1940. En effet, Delbrück mettait déjà en avant le fait que les attracteurs modélisaient les fonctions biologiques (Delbrück, 1950), ce qui s'est vu largement confirmé par la suite. L'exemple d'application qui en est certainement le plus représentatif est celui issu des travaux de Thieffry et Thomas sur le bactériophage lambda dans lesquels ils sont parvenus à mettre en évidence le passage d'un phase de lyse à une phase de lysogénie (Thieffry, Thomas, 1995). Les attracteurs jouent donc un rôle fondamental car ils expriment la capacité des réseaux biologiques à produire certaines fonctions. En conséquence, nombreuses sont les recherches fondamentales qui se sont concentrées, et se concentrent encore, sur les comportements asymptotiques des réseaux d'automates.

Nous évoquons à présent les théorèmes qui tiennent lieu de références dans ce domaine et qui mettent en avant les architectures cycliques comme motrices de la complexité du vivant. Le premier est de Robert et précise le fait que tout réseau d'automates dont le graphe d'interaction est acyclique possède un comportement dynamique trivial, à savoir qu'il admet comme unique attracteur une unique configuration stable (Robert, 1986 ; Robert, 1995). De ce théorème découle naturellement que les réseaux biologiques possèdent nécessairement un cycle dans leur architecture pour garantir la complexité et la richesse comportementale inhérentes aux systèmes biologiques. Les suivants sont connus comme les deux conjectures de Thomas, que ce dernier a introduites dans (Thomas, 1981) et qui ont été depuis démontrées dans le contexte discret (Richard, Comet, 2007 ; Remy, Ruet et al., 2008 ; Richard, 2010). Ils mettent en avant le rôle des cycles positifs et négatifs dans les réseaux, un cycle positif (resp. négatif) étant défini comme un cycle, au sens de la théorie des graphes, possédant un nombre pair (resp. impair) d'arêtes négatives (i.e. une arête marquant l'action inhibitrice d'un automate sur un autre). Ces théorèmes sont généralement énoncés comme suit, sous l'hypothèse du mode de mise à jour asynchrone : (1) la présence d'un cycle positif dans le graphe d'interaction d'un réseau d'automates est nécessaire à sa multi-stationnarité, c'est-à-dire à l'existence de plusieurs configurations stables dans son graphe de transition; (2) la présence d'un cycle négatif est nécessaire à l'existence d'une oscillation stable.

Pour conclure cette partie illustrant le point de vue théorique des recherches entreprises en modélisation des réseaux biologiques, nous présentons des grandes thématiques de recherche interdisciplinaires qui nous semblent parmi les plus pertinentes en modélisation des réseaux, et qui ont toutes en tâche de fond la volonté d'établir des liens entre les caractéristiques architecturelles et dynamiques des réseaux d'automates :

- *Trajectoires et asymptotes* : ce thème a pour objectif d'aboutir à des caractérisations des comportements des réseaux, qu'ils soient trajectoriels ou asymptotiques. Ce que l'on entend ici par comportement trajectoriel concerne la partie des graphes transitions qui sont hors des attracteurs. Bien que nous n'ayons pas évoqué ces comportements dans ce qui précède, ils s'avèrent particulièrement utiles pour comprendre les systèmes biologiques. En particulier, de nombreux problèmes sont ouverts quant aux temps d'attraction (c'est-à-dire le nombre d'étapes de temps nécessaire pour que l'ensemble des configurations atteignent de manière certaines leur(s) attracteur(s)), qui du point de vue de la biologie permettent d'étudier par exemple à quelle vitesse les réseaux de régulation génétique se spécialisent rapidement vers des fonctions particulières. Certaines études contemporaines fournissent des résultats à ce propos mais ils sont généralement valides pour des familles spécifiques de réseaux d'automates (Orponen, 1997 ; Goles, Hernandez, 2000 ; Noual, Regnault et al., 2012). D'autres études ont été menées sur le concept de bassins d'attraction, le bassin d'attraction associé à un attracteur correspondant à l'ensemble des configurations transitoires attirées par cet attracteur (Wuensche, 1998 ; Demongeot, Goles et al., 2010). Ces bassins peuvent fournir des informations intéressantes quant à la vraisemblance qu'un système a de se spécialiser vers telles ou telles fonctions, ce qui est a priori pertinent en biologie. Par ailleurs, nous avons précédemment évoqué l'importance des attracteurs dans les réseaux. Au delà de la caractérisation de leur nature, sujet des travaux de de Thomas et Robert évoqués plus haut, ou encore de ceux de Goles (Goles, Olivos, 1981 ; Aracena, Demongeot et al. 2004), de plus en plus de recherches ont mis en évidence l'importance de parvenir à les compter, qu'on rapproche aisément de la combinatoire, domaine au cœur de la science informatique (Sontag, Veliz-Cuba et al., 2008 ; Richard, 2008 ; Demongeot, Noual et al., 2012).
- *Modularité* : l'étude de la modularité des réseaux de régulation biologique se trouve au cœur de recherches de plus en plus nombreuses. Comprendre à quel point les réseaux sont modulaires est une problématique centrale en bio-informatique des réseaux pour plusieurs raisons, qu'elles soient de nature informatique ou biologique. Bien sûr, du point de vue informatique, l'étude des comportements des réseaux soulève le problème fondamental de l'explosion combinatoire. En effet, même si l'on considère l'abstraction la plus élevée en vue de modéliser les réseaux de régulation en s'attachant à étudier des réseaux d'automates booléens, la taille de leurs graphes de transition augmente exponentiellement en fonction de leur taille. En effet, le comportement dynamique d'un réseau d'automates de taille n se représente par un graphe de transition dont le nombre de sommets est 2^n. C'est notamment pour cette raison que la littérature s'attache principalement à étudier des réseaux de petite taille. Or, l'étude de la modularité des réseaux a en partie pour vocation à réduire ce facteur exponentiel. Du point de vue de la biologie maintenant, l'idée générale est de réussir à mettre en évidence dans les réseaux de régulation des sous-motifs qui possèdent des fonctionnalités propres, relativement indépendants du reste, qui reste l'une des problèmes phares en modélisation. Plusieurs méthodes ont jusqu'à présent été mises en avant dans ce contexte pour caractériser la modularité des réseaux, qui se fondent sur des approches intrinsèquement éloignées. La première famille de méthodes vise à extraire des graphes d'interaction des sous-motifs (de petite taille si possible) largement présents dans la réalité, au moyen de techniques provenant principalement des statistiques (Spirin, Mirny, 2003 ; Alon, 2003). *A contrario,* la deuxième famille prend comme objets d'étude les graphes de transition des réseaux et vise à découvrir les modules de manière purement dynamique (Thieffry, Romero, 1999 ; Segal, Shapira et al., 2003). Enfin, des recherches récentes et prometteuses ont mis en avant une approche originale conjuguant les aspects statiques et dynamiques des réseaux

(Bernot, Tahi, 2009 ; Siebert, 2009 ; Delaplace, Klaudel et al. 2012), qui mérite que des efforts soient poursuivis dans le même sens.

- *Robustesse* : La robustesse des réseaux est la dernière thématique que nous mettons en avant dans le cadre de la modélisation des réseaux de régulation. Informellement, les recherches menées dans ce cadre visent à comprendre comment les réseaux réagissent à des perturbations qui peuvent être de différentes natures. En biologie, plusieurs types de robustesse méritent d'être étudiés plus avant. Nous précisons dans la suite quatre d'entre eux qui nous paraissent les plus pertinents :
    - la *robustesse comportementale* : les études sur ce type de robustesse visent à étudier comment les systèmes réagissent à des changements de conditions initiales. L'intérêt des études relatives à ce type de robustesse est, selon les connaissances qui en découlent, qu'elles permettent d'éviter de prendre en considération l'intégralité du comportement des systèmes, et donc de réduire là encore le problème lié à l'explosion combinatoire sous-jacente ;
    - la *robustesse architecturelle* : cette classe de robustesse vise à analyser l'action de variations temporaires ou définitives de l'architecture des modèles sur leurs comportements. En biologie, ce type de robustesse est particulièrement adapté pour mieux comprendre l'effet de certains éléments sur les régulations. Un exemple est celui des micro-A.R.N. dont l'action d'inhibition post-transcriptionnelle peut être assimilée à la rupture d'interactions existant entre des gènes au sein d'un réseau de régulation ;
    - la *robustesse environnementale* : ici, l'objectif est de comprendre si le comportement des systèmes est conservé quand ils sont sujets à des perturbations qui ne relèvent pas de variations internes mais de contraintes exercées depuis l'extérieur : ce type de relation entre les systèmes et leur contexte a été largement traité dans toutes les disciplines, et a récemment fait l'objet d'un ouvrage en biologie sur les interactions réciproques entre les bactéries et leur milieu (Guespin-Michel, 2011).
    - la *robustesse structurelle* : ce dernier type de robustesse consiste à analyser l'influence de transformations fonctionnelles (à savoir associées au comportement). Des éléments fondateurs ont été fournis par Thom dans (Thom, 1972). Dans le contexte qui nous occupe, toutes les études menées par Robert et les travaux qui en ont découlé sur les influences des modes de mises à jour, s'apparentent à ce type de robustesse (Robert, 1986 ; Robert, 1995). Des études de ce type pourraient notamment permettre au niveau biologique d'en apprendre plus sur la dynamique chromatinienne, les transformations de la chromatine agissant directement sur l'ordonnancement des régulations dans le temps.

3.2 Modélisation appliquée

En biologie, le plus souvent, nous proposons déjà un « modèle » simple, consistant à établir des relations entre molécules ou processus biologiques et à les intégrer dans un ensemble expérimental défini (par exemple : le modèle de la sécrétion de l'enzyme *E* par des cellules *C* dans telles conditions conduit à activer la voie *V*, le tout étant attesté par l'augmentation de la protéine *P*). Toutefois, une des caractéristiques de nos modèles est de négliger certains éléments qui nous paraissent non essentiels au mécanisme considéré (comme on néglige, par exemple, les conditions biomécaniques alors qu'elles peuvent influencer les comportements cellulaires autant que certaines molécules « effectrices »). *A contrario*, parfois, nous surestimons l'importance d'autres éléments. Nous « contextualisons » aussi les relations entre éléments, ce qui mène à ne pas pouvoir tirer de conclusions autres que celles qui se réfèrent à notre situation expérimentale précise alors que, pour atteindre une vision globale des évènements, il faut réduire le nombre de paramètres. En fait, le choix de la modélisation dépend d'une part, de la taille du réseau biologique considéré et, d'autre part, du réseau scientifique (biologistes-modélisateurs) constitué. Pour illustrer ce point,

reprenons à notre compte les termes employés dans (Herzel, Blüthgen, 2008) : « *Pour de petits sous-systèmes, les modèles dynamiques peuvent être adaptés directement aux données expérimentales. Mais malgré des microaarrays de plus en plus performants, ainsi que les données obtenues par protéomique ou imagerie, les plus grands systèmes, comme la machinerie du cycle cellulaire, ne peuvent pas encore être décrits par des modèles mathématiques. Dans ces cas, les approches statistiques comme la P.C.A., les réseaux bayésiens et la logique floue, peuvent aider à extraire des informations utiles.*"

Pour notre part, nous avons décrit un système biologique relativement simple : autour des cellules cancéreuses une molécule, PAI 1, peut être sécrétée, inhibée, activée, internalisée, utilisée… (cf. Figure 7). Cette molécule, lorsqu'elle est liée à la matrice extra-cellulaire et active, est capable d'influencer le comportement de cellules cancéreuses en promouvant la transition mésenchymo-amaeboide. Les cellules cancéreuses qui ont passé cette transition morphologique, changent de comportement et notamment de mode de migration puisqu'elle adoptent alors la migration amaeboide. Or, ce type de migration est considéré comme « le » type de migration métastatique (cf. chapitre {« Dynamique du micro-environnement cellulaire »}). L'une des questions qui revient alors est la suivante : comment passer d'une observation *in vitro*, à l'énoncé d'une loi de comportement valable *in vivo* ? C'est justement pour répondre à ce type de question que la modélisation peut s'avérer utile mais cela se fait en plusieurs étapes.

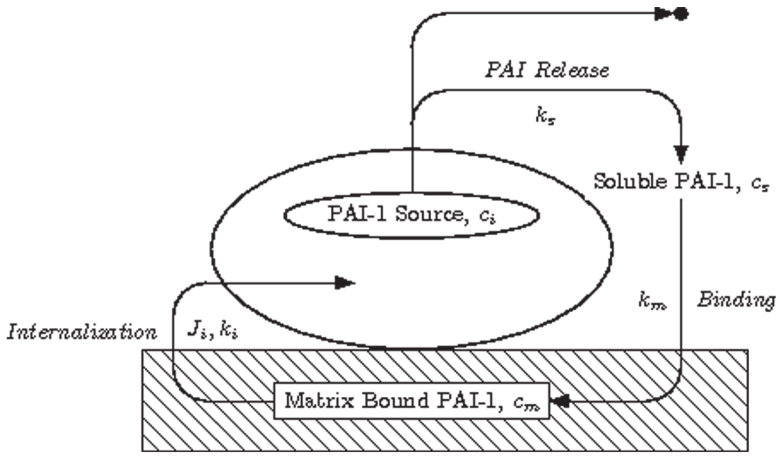

Fig. 7 : Flux et évolution de trois conformères de PAI 1 : produit dans la cellule, il est secrété sous forme soluble inactive ; s'il se lie à la matrice (via la Vitronectine) il devient PAI 1 matriciel actif, puis, s'il est « utilisé » par la cellule, il est internalisé et détruit (image issue de (Malo, Cartier-Michaud, et al. 2010)).

En l'occurrence, dans le cadre qui nous intéresse, nous avions montré expérimentalement qu'une grande concentration du conformère actif PAI 1 matriciel favorisait la transition mésenchymo-amaeboide de cellules cancéreuses invasives. Ceci suggérait que la « rencontre » d'une cellule cancéreuse avec ces conditions (PAI 1 matriciel actif) pouvait représenter une condition favorable à l'échappement métastatique. Il nous fallait *alors vérifier si le scénario était plausible* et *comment se déposait une molécule secrété par une tumeur en croissance*. Dans un modèle de réaction-diffusion, nous avons figuré les niveaux de production de cette molécule, d'utilisation et de contrôle au cours du processus métastatique (Malo, Cartier-Michaud et al., 2010 ; Cartier-Michaud, Malo et al., 2012). De là, pour modéliser le dépôt de PAI 1 autour d'une tumeur en croissance, nous avons implanté un simulateur du modèle fondé sur la théorie des automates cellulaires.

Les résultats de simulation ainsi obtenus mettent en exergue qu'une situation minimale, c'est-à-dire la croissance stochastique de cellules cancéreuses couplée à la diffusion et au dépôt de PAI-1 secrété (équivalent pour n'importe quelle autre protéine sécrétée) dans la matrice extra-cellulaire disponible, est suffisante pour rendre compte de la distribution très hétérogène de PAI-1 matriciel à la périphérie de la

tumeur. On y voit également que les pics de concentration correspondent à une « géographie » de la tumeur plus invaginée, ce qui est en adéquation avec les observations cliniques.

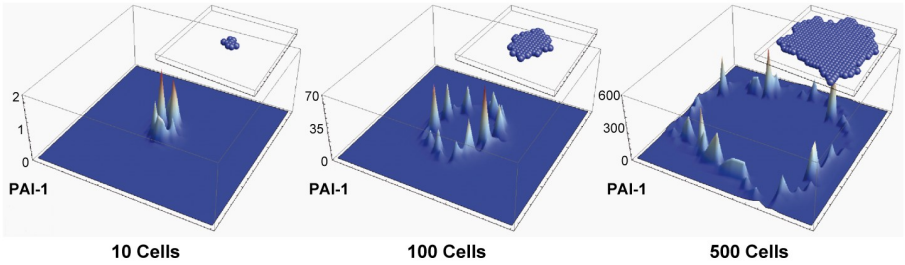

Fig. 8 : Résultats de simulations du modèle de réaction-diffusion obtenus à partir du simulateur fondé sur la théorie des automates cellulaires (image issue de (Cartier-Michaud, Malo et al., 2012)).

Pour aller plus loin, un autre simulateur a été implanté en se fondant sur le paradigme de la simulation multi-agents. Sans qu'il soit ici question du paradigme, ce simulateur prend en considération plus de paramètres que le précédent, comme la nécrose au centre de la tumeur et la protéolyse (due à l'activité enzymatique) à la périphérie de la tumeur. De fait, il est ainsi plus complet et permet une prise en compte plus réaliste du modèle. Les résultats obtenus sur cette base (cf. Figure 9), plus fins, reproduisent notamment ce qui a été décrit sur la structure des tumeurs, ainsi que la distribution hétérogène de PAI-1 matriciel autour de la tumeur (Umeda, Eguchi et al., 1997).

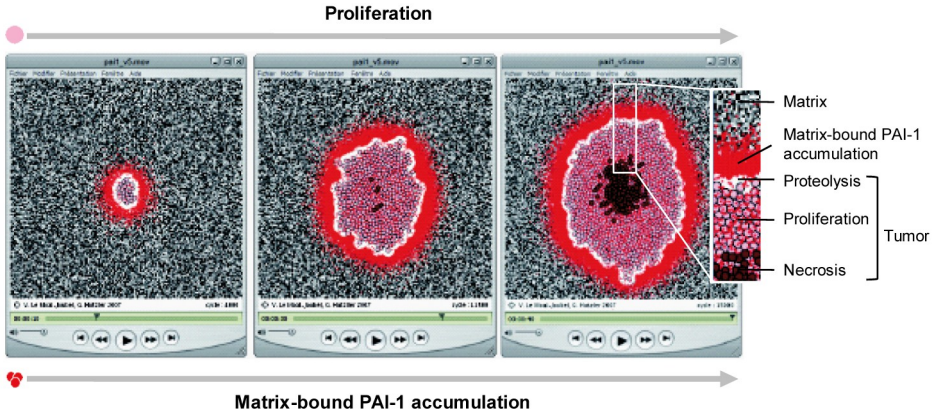

Fig 9 : Résultats de simulations du modèle de réaction-diffusion obtenus à partir du simulateur fondé sur le paradigme multi-agents montrant la prolifération de PAI-1 au cours du temps (image issue de (Cartier-Michaud, Malo et al., 2012)).

Ainsi, une cellule donnée à la périphérie d'une tumeur peut théoriquement rencontrer de hautes concentrations de PAI-1 matriciel et l'utiliser, comme nous l'avons montré in vitro, entraînant une transition morphologique mésenchymo-amaeboide, et l'adoption de la migration amaeboide adaptée. Les deux types de simulations suggèrent le mécanisme suivant : toutes les cellules de la tumeur produisent du PAI-1 matriciel qui est utilisé par une petite fraction d'entre elles. Ceci est favorisé par la morphologie locale de la tumeur (les invaginations qui induisent des dépôts plus importants). Ils suggèrent également que le caractère rare de l'échappement métastatique trouve son origine dans le contrôle de l'internalisation de PAI-1, de sa signalisation (Cartier-Michaud, Malo et al., 2012) et dans sa capacité à promouvoir le passage entre les états mésenchymateux et amaeboides.

Aussi, la modélisation des effets possibles de changements de concentration de PAI-1 matriciel, dans un cycle suivi par une cellule, indique que ces changements contrôlent le passage entre deux états particuliers

associés pour l'un au comportement mésenchymateux et, pour l'autre, au comportement amaeboide. En effet, une cellule passe d'un comportement mésenchymateux à un comportement amaeboide lorsqu'elle rencontre des concentrations élevées de PAI-1. Lorsqu'elle devient amaeboide, elle peut adopter la migration de même type ; dès lors elle quittera progressivement le « spot » de haute concentration de PAI-1 pour des régions de plus basses concentrations de PAI-1. Dans ce cas elle subira la transition morphologique inverse et retrouvera son caractère mésenchymateux, et le comportement ad hoc. On comprend alors, que les transitions d'un état à l'autre sont des situations très instables alors que les états amaeboide ou mésenchymateux qui ne doivent leur instabilité théorique qu'à un changement de la concentration de PAI-1 peuvent être considérés comme des situations d'équilibre. En réalité, cettre dynamique suit un cycle hystérétique, bien connu en physiologie.

Les différents modèles informatiques implantant le modèle mathématique de réaction-diffusion, qu'il soient simples ou plus complexes, suggèrent tous que des concentrations hétérogènes de molécules matricielles existent à la périphérie des tumeurs. Les cellules qui se détachent de la tumeur peuvent donc rencontrer la concentration de PAI 1 adéquate pour une transition morphologique. Dès lors, elles peuvent adopter une migration amaeboide et participer à l'échappement métastatique. Ces modèles suggèrent également que le micro-environnement représente une cible thérapeutique, en particulier le « conformère » de PAI-1 matriciel et actif. Ainsi, les modèles informatiques, associés aux résultats expérimentaux *in vitro*, permettent de proposer le blocage matriciel du PAI-1 actif comme une voie importante d'approche du processus métastatique. Bien entendu, il faudra confronter les résultats théoriques de ce blocage avec les effets prédictibles sur d'autres systèmes comme la protéolyse, la fibrinolyse ou l'inflammation… Certes, il n'y a rien de magique, mais la modélisation permet lorsqu'elle est associée à des résultats expérimentaux, d' approcher un peu mieux la réalité complexe d'un processus biologique.

4. Conclusion

Modéliser au plus proche des préoccupations qui relèvent de la biologie nécessite avant tout un dialogue interdisciplinaire. Ne pas attendre ni proposer la « solution à tout », mais plutôt, valider une hypothèse, mieux comprendre les règles de base et l'essence des systèmes, aboutir à une simulation qui fait apparaître un comportement inédit, une courbe de comportement, ou encore, inscrire dans une logique inattendue, les résultats expérimentaux apparemment dissonants : telle est notre humble expérience...

Dans *Le chemin de l'espérance*, en 2011, Hessel et Morin observent qu'une réorganisation du savoir est en cours et plaident, évidement, « pour l'interdisciplinarité en attendant qu'on reconnaisse la transdisciplinarité..., dans le cadre de la pensée complexe". Nous joindrions bien volontiers nos espoirs aux leurs tant il est impossible, dans la modélisation informatique pour la biologie, de démêler les approches ; même si chaque discipline garde ses méthodes propres, la culture qui permet la modélisation, est celle qui devient commune.

Bibliographie

U. Alon. Biological Networks; The Tinkerer as an Engineer. *Science*, vol. 301, p. 1866-1867, 2003.

U. Alon. Network Motifs: Theory and Experimental Approaches, *Nature Reviews Genetics*, vol. 8, p. 450-461, 2007.

S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman. Basic Local Alignment Search Tool. *Journal of Molecular Biology*, vol. 215, p. 403-410, 1990.

J. Aracena, J. Demongeot, E. Goles. Fixed Points and Maximal Independent Sets in AND-OR Networks. *Discrete*

Applied Mathematics, vol. 138, p. 277-288, 2004.

C. Berge. *Théorie des graphes et ses applications*, Dunod, 1958.

G. Bernot, F. Tahi. Behaviour Preservation of a Biological Regulatory Network When Embedded Into a Larger Network. *Fundamenta Informaticae*, vol. 91, p. 463-485, 2009.

L. K. Billings, J. C. Florez. The Genetics of Type 2 Diabetes: What Have We Learned from GWAS?, *Annals of the New York Academy of Science*, vol. 1212, p. 59-77, 2010.

C. M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, 1995.

P. E. Bourne, H. Weissig (éds.). *Structural Bioinformatics*, Wiley, 2003.

I. Brierley, R. J. C. Gilbert, S. Penell. RNA Pseudoknots and the Regulation of Protein Synthesis. *Biochemical Society Transactions*, vol. 36, p. 684-689, 2008.

M. C. Brown, C. E. Turner. Paxillin: Adapting to Change. Physiological Reviews, vol. 84, p. 1315-1339, 2004.

C. Burge, S. Karlin. Prediction of Complete Gene Structures in Human Genomic DNA. *Journal of Molecular Biology*, vol. 268, p. 78-94, 1997.

W. S. Bush, J. H. Moore. Genome-Wide Association Studies. Chap. 11 de Translational Bioinformatics. *PLoS Computational Biology*, vol. 8, e1002822, 2012.

A. Cartier-Michaud, M. Malo, C. Charrière-Bertrand, G. Gadea, C. Anguille, A. Supiramaniam, A. Lesne, F. Delaplace, G. Hutzler, P. Roux, D. A. Lawrence, G. Barlovatz-Meimon. Matrix-Bound PAI-1 Supports Cell Blebbing via RhoA/ROCK1 Signaling. *PloS One*, vol. 7, p. e32204, 2012.

O. Carton. *Langages formels - Calculabilité et complexité*, Vuibert, 2008.

J. Y. Chen, S. Lonardi. *Biological Data Mining*, CRC Press, 2009.

A. Church. *The Calculi of Lambda-Conversion*, Princeton University Press, 1941.

E. F. Codd. A Relational Model of Data for Large Shared Data Banks. *Communications of the ACM*, vol. 13, p. 377-387, 1970.

T. H. Cormen, C. E. Leiserson, R. L. Rivest, C. Stein. *Introduction à l'algorithmique*, Dunod, 1994.

M. Crochemore, C. Hancart, T. Lecroq. *Algorithmique du texte*, Vuibert, 2001.

V. Cutello, G. Narzisi, G. Nicosia. A Multi-Objective Evolutionary Approach to the Protein Structure Prediction Problem. *Journal of the Royal Society Interface*, vol. 3, p. 139-151, 2006.

C. Darwin. *Charles Darwin's Notebooks (1836-1844)*. P. H. Barrett, P. J. Gautrey, S. Herbert, D. Kohn, S. Smith (éd.), Cornell University Press, 1987.

O. Delaneau, J. Marchini, J.-F. Zagury. A Linear Complexity Phasing Method for Thousands of Genomes. *Nature Methods*, vol. 9, p. 179-181, 2012.

F. Delaplace, H. Klaudel, T. Melliti, S. Sené. Analysis of Modular Organisation of Interaction Networks Based on Asymptotic Dynamics. *Proceedings of CMSB*, vol. 7605 de *Lecture Notes in Computer Science*, p. 148-165, Springer, 2012.

M. Delbrück (éd.). *Viruses*, California Institute of Technology, 1950.

J. Demongeot, E. Goles, M. Morvan, M. Noual, S. Sené. Attraction basins as gauges of the robustness against boundary conditions in biological complex systems. *PLoS One*, vol. 5, p. e11793, 2010.

J. Demongeot, M. Noual, S. Sené. Combinatorics of Boolean Automata Circuits Dynamics. *Discrete Applied Mathematics*, vol. 160, p. 398-415, 2012.

A. Dupressoir, T. Heidmann. Les syncytines : des protéines d'enveloppe rétrovirales capturées au profit du développement placentaire. *Médecine/Sciences*, vol. 27, p. 163-169, 2011.

A. Dupressoir, C. Vernochet, O. Bawa, F. Harper, G. Pierron, P. Opolon, T. Heidmann. Syncytin-A knockout mice demonstrate the critical role in placentation of a fusogenic, endogenous retrovirus-derived, envelope gene. *Proceedings of the National Academy of Sciences of the USA*, vol. 106, p. 12127-12132, 2009.

R. Durbin, S. Eddy, A. Krogh, G. Mitchison. *Biological Sequence Analysis*, Cambridge University Press, 1998.

L. Euler. Solutio problematis ad geometriam situs pertinentis. *Commentarii Academiae Scientiarum Petropolitanae*,

vol. 8, p. 128-140, 1741.

S. Feferman, J. W. Dawson, S. C. Kleene, G. H. Moore, R. M. Solovay et J. van Heijenoort (eds.). *Kurt Gödel Collected Works, Volume I – Publications 1929-1936*. Oxford University Press, 1986.

J. Felsenstein. Cases in Which Parsimony and Compatibility Methods will be Positively Misleading. *Systematic Zoology*, vol. 27, p. 401-410, 1978.

J. Felsenstein. Inferring Phylogenies from Protein Sequences by Parsimony, Distance, and Likelihood Methods. *Methods in Enzymology*, vol. 266, p. 418-427, 1996.

W. M. Fitch. Toward Defining the Course of Evolution: Minimum Change for a Specified Tree Topology. *Systematic Zoology*, vol. 20, p. 406-416, 1971.

F. L. G. Fredge. *Die Grundlagen der Arithmetik* (Les fondements de l'arithmétique), W. Koebner, 1884.

O. Gascuel (éd.). *Mathematics of Evolution & Phylogeny*. Oxford Universty Press, 2005.

O. Gascuel, M. A. Steel (éds.). *Reconstructing Evolution: New Mathematical and Computational Advances*. Oxford University Press, 2007.

J. Guespin-Michel. *Les bactéries, leur monde et nous : vers une biologie intégrative et dynamique*, Dunod, 2011.

S. Guindon, J. F. Dufayard, V. Lefort, M. Anisimova, W. Hordijk, O. Gascuel. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Systematic Biology*, vol. 59, p. 307-321, 2010.

S. Guindon, O. Gascuel. A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood. *Systematic Biology*, vol. 52, p. 696-704, 2003.

E. Goles, G. Hernandez. Dynamical Behavior of Kauffman Networks with AND-OR Gates. *Journal of Biological Systems*, vol. 8, p. 151-175, 2000.

E. Goles, J. Olivos. The Convergence of Symmetric Threshold Automata. *Information and Control*, vol. 51, p. 98-104, 1981.

M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten. The WEKA Data Mining Software: An Update. *ACM SIGKDD Explorations Newsletter*, vol. 11, p. 10-18, 2009.

H. Herzel, N. Blüthgen. Mathematical Models in Mammalian Cell Biology. *Genome Biology*, vol. 9, article 316, 2008.

P. Hogeweg, B. Hesper. Interactive Instruction on Population Interactions. *Computers in Biology and Medicine*, vol. 8, p. 319-327, 1978.

J. E. Hopcroft, J. D. Ullman. *Introduction to Automata Theory, Languages, and Computation*, Addison-Wesley Publishing, 1979.

F. Jacob. *Les bactéries lysogènes et la notion de provirus*. Masson, 1954.

F. Jacob, J. Monod. Genetic Regulatory Mechanisms in the Synthesis of Proteins. *Journal of Molecular Biology*, vol. 3, p. 318-356, 1961.

F. Jacob, D. Perrin, C. Sanchez, J. Monod. L'opéron : groupe de gènes à expression coordonnée par un opérateur. *Comptes rendus hebdomadaires de l'Académie des Sciences*, vol. 250, p. 1727-1729, 1960.

H. F. Jordan, G. Alaghband. *Fundamentals of Parallel Processing*, Prentice Hall, 2002.

S. A. Kauffman. Metabolic stability and epigenesis in randomly constructed genetic nets. *Journal of Theoretical Biology*, vol. 22, p. 437-467, 1969.

N. Kim, K. N. Fuhr, T. Schlick. Graph Applications to RNA Structure and Function. Chap. 3 de *Biophysics of RNA Foldings*, Springer, 2013.

W. Krauth. Introduction to Monte Carlo Algorithms, arXiv:cond-mat/9612186, 1996.

K. R. Legate, E. Montañez, O. Kudlacek, R. Füssler. ILK, PINCH and parvin: the tIPP of Integrin Signalling. *Nature Reviews Molecular Cell Biology*, vol. 7, p. 20-31, 2006.

P. O. Lewis. A Genetic Algorithm for Maximum-Likelihood Phylogeny Inference Using Nucleotide Sequence Data. *Molecular Biology and Evolution*, vol. 15, p. 277-283, 1998.

L. Liberti, C. Lavor, N. Maculan. A Branch-and-Prune Algorithm for the Molecular Distance Geometry Problem. *International Transactions in Operational Research*, vol. 15, p. 1–17, 2008.

S. Limou, J.-F. Zagury. Immunogenetics: Genome-Wide Association of Non-Progressive HIV and Viral Load Control: *HLA* Genes and Beyond. *Frontiers in Immunology*, vol. 4, article 118, 2013.

D. J. Lipman, W. R. Pearson. Rapid and Sensitive Protein Similarity Searches. *Science*, vol. 227, p. 1435-1441, 1985.

S. H. Lo. Focal Adhesions: What's New Inside. *Developmental Biology*, vol. 294, p. 280-291, 2006.

C.-M. Lo, H.-B. Wang, M. Dembo, Y.-L. Wang. Cell Movement Is Guided by the Rigidity of the Substrate. *Biophysical Journal*, vol. 79, p. 144-152, 2000.

P. Lopez, D. Casane, H. Philippe. Phylogénie et évolution moléculaires. *Médecine/Sciences*, vol. 18, p. 1146-1154, 2002.

A. Lovelace, B. A. Toole (éd.). *Ada, the Enchantress of Numbers: A Selection from the Letters of Lord Byron's Daughter, and her Description of the First Computer*, Strawberry, 1992.

A. Lwoff. Factors Influencing the Evolution of Viral Diseases at the Cellular Level and in the Organism. *Bacteriology Reviews*, vol. 23, p. 109-124, 1959.

A. Lwoff, L. Siminovitch, N. Kjeldgaard. Induction de la production de bactériophages chez une bactérie lysogène. *Annales de l'Institut Pasteur*, vol. 79, p. 815-859, 1950.

M. Malo, A. Cartier-Michaud, E. Fabre-Guillevin, G. Hutzler, F. Delaplace, G. Barlovatz-Meimon, A. Lesne. When a Collective Outcome Triggers a Rare Individual Event: A Mode of Metastatic Process in a Cell Population. *Mathematical Population Studies*, vol. 17, p. 136-165, 2010.

A. M. Maxam et W. Gilbert. A New Method for Sequencing DNA. *Proceedings of the National Academy of Sciences of the USA*, vol. 74, p. 560-564, 1977.

W. S. McCulloch et W. Pitts. A Logical Calculus of the Ideas Immanent in Nervous Activity. *Bulletin of Mathematical Biophysics*, vol. 5, p. 115-133, 1943.

J. Monod. *Recherches sur la croissance des cultures bactériennes*, Hermann, 1958.

J. Monod, E. Wollman. Inhibition de la croissance et de l'adaptation enzymatique chez les bactéries infectées par le bactériophage. *Annales de l'Institut Pasteur*, vol. 73, p. 937-956, 1947.

J. J. Moré, Z. Wu. Distance Geometry Optimization for Protein Structures. *Journal of Global Optimization*, vol. 15, p. 219-234, 1999.

R. Mott. EST_GENOME: A Program to Align Spliced DNA Sequences to Unspliced Genomic DNA. *CABIOS Applications Notes*, vol. 13, p. 477-478, 1997.

F. Murtagh. Complexities of Hierarchic Clustering Algorithms: the state of the art. *Computational Statistics Quarterly*, vol. 1, p. 101-113, 1984.

Myocardial Infarction Genetics Consortium. Genome-Wide Association of Early-Onset Myocardial Infarction With Single Nucleotide Polymorphisms and Copy Number Variants. *Nature Genetics*, vol. 41, p. 334-341, 2009.

M. Noual. *Updating Automata Networks*. Thèse de doctorat, École normale supérieure de Lyon, 2012.

M. Noual, D. Regnault, S. Sené. Boolean networks synchronism sensitivity and XOR circulant networks convergence time. *Full Papers Proceedings of AUTOMATA & JAC*, vol. 90 des *Electronic Proceedings in Theoretical Computer Science*, p. 37-52, Open Publishing Association, 2012.

M. Noual, D. Regnault, S. Sené. About Non-Monotony in Boolean Automata Networks. *Theoretical Computer Science*, vol. 504, p. 12-25, 2013.

P. Orponen. Computing with Truly Asynchronous Threshold Logic Networks. *Theoretical Computer Science*, vol. 174, p. 123-136, 1997.

R. D. M. Page, E. C. Holmes. *Molecular Evolution: A Phylogenetic Approach*, Blackwell Publishing Company, 1998.

W. R. Pearson, D. J. Lipman. Improved Tools for Biological Sequence Comparison. *Proceedings of the National Academy of Sciences of the USA*, vol. 85, p. 2444-2448, 1988.

É. Remy, P. Ruet, D. Thieffry. Graphic Requirement for Multistability and Attractive Cycles in a Boolean Dynamical Framework. *Advances in Applied Mathematics*, vol. 41, p. 335-350, 2008.

A. Richard. Positive Circuits and Maximal Number of Fixed Points in Discrete Dynamical Systems. *Discrete Applied Mathematics*, vol. 157, p. 3281-3288, 2009.

A. Richard. Negative Circuits and Sustained Oscillations in Asynchronous Automata Networks. *Advances in Applied Mathematics*, vol. 44, p. 378-392, 2010.

A. Richard, J.-P. Comet. Necessary Conditions for Multistationarity in Discrete Dynamical Systems. *Discrete Applied Mathematics*, 155(18):2403-2413, 2007.

F. Robert. *Discrete Iterations: A Metric Study*, Springer, 1986.

F. Robert. *Les systèmes dynamiques discrets*, Springer, 1995.

N. Saitou, M. Nei. The Neighbor-Joining Method: A New Method for Reconstructing Phylogenetic Trees. *Molecular Biology and Evolution*, vol. 4, p.406-425, 1987.

F. Sanger, G. M. Air, B. G. Barrell, N. L. Brown, A. R. Coulson, C. A. Fiddes, C. A. Hutchison, P. M. Slocombe et M. Smith. Nucleotide Sequence of Bacteriophage Phi X174 DNA. *Nature*, vol. 265, p. 687-695, 1977.

G. Saporta. *Probabilités, analyse des données et statistiques*, Éditions Technip, 2006.

E. Segal, M. Shapira, A. Regev, D. Pe'er, S. Botstein, D. Koller, N. Friedman. Module Networks: Identifying Regulatory Modules and their Condition-Specific Regulators from Gene Expression Data. *Nature Genetics*, vol. 34, p. 166-176, 2003.

C. Semple, M. A. Steel. *Phylogenetics*, Oxford University Press, 2003.

S. Sené. *La bio-informatique des réseaux d'automates*. Thèse d'habilitation à diriger des recherches, Université d'Évry - val d'Essonne, 2012.

H. Siebert. Dynamical and Structural Modularity of Discrete Regulatory Networks. *Proceedings of CompMod*, vol. 6 de *Electronic Proceedings in Theoretical Computer Science*, p. 109-124, Open Publishing Association, 2009.

N. J. A. Sloane et A. D. Wyner. *Claude Elwood Shannon: Collected Works*, IEEE, 1993.

R. Sokal, C. Michener. A Statistical Method for Evaluating Systematic Relationships. *University of Kansas Science Bulletin*, vol. 38, p. 1409-1438, 1958.

E. Sontag, A. Veliz-Cuba, R. Laubenbacher, A. S. Jarrah. The Effect of Negative Feedback Loops on the Dynamics of Boolean Networks. *Journal of Biophysics*, vol. 95, p. 518-526, 2008.

V. Spirin, L. A. Mirny. Protein Complexes and Functional Modules in Molecular Networks. *Proceedings of the National Academy of Sciences of the USA*, vol. 100, p. 12123-12128, 2003.

R. Staden. A Strategy of DNA Sequencing Employing Computer Programs. *Nucleic Acids Research*, vol. 6, p. 2601-2610, 1979.

D. Thieffry, D. Romero. The Modularity of Biological Regulatory Networks. *Biosystems*, vol. 50, p. 49-59, 1999.

D. Thieffry, R. Thomas. Dynamical behaviour of biological regulatory networks – II. Immunity control in bacteriophage lambda. *Bulletin of Mathematical Biology*, vol. 57, p.277-297, 1995.

R. Thom. *Stabilité structurelle et morphogenèse : vers une théorie générale des modèles*, Interéditions, 1972.

R. Thomas. Boolean formalisation of genetic control circuits. *Journal of Theoretical Biology*, vol. 42, p. 565-583, 1973.

R. Thomas. On the relation between the logical structure of systems and their ability to generate multiple steady states or sustained oscillations. *Numerical methods in the study of critical phenomena*, vol. 9 de *Springer Series in Synergetics*, p. 180-193, Springer, 1981.

R. Thomas. Regulatory Networks Seen as Asynchronous Automata: A Logical Description. *Journal of Theoretical Biology*, vol. 153, p. 1-23, 1991.

Transparency Market Research. *Bioinformatics Market - Global Industry Size, Market Share, Trends, Analysis and Forecast, 2012-2018*, 2012.

A. M. Turing. On Computable Numbers, with an Application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society*, vol. 2, p. 230-265, 1936.

T. Umeda, Y. Eguchi, K. Okino, M. Kodama, T. Hattori. Cellular Localization of Urokinase-Type Plasminogen Activator, Its Inhibitors, and Their mRNAs in Breast Cancer Tissues. *The Journal of Pathology*, vol. 183, p. 388-397, 1997.

K. Wang, W.-D. Li, C. K. Zhang, Z. Wang, J. T. Glessner, S. F. A. Grant, H. Zhao, H. Hakonarson, R. Arlen Price. A Genome-Wide Association Study on Obesity and Obesity-Related Traits. PLoS One, vol. 6, e18939, 2011.

R. H. Waterston, E. S. Lander, J. E. Sulston. On the Sequencing of the Human Genome. *Proceedings of the National Academy of Sciences of the USA*, vol. 99, p. 3712-3716, 2002.

J. L. Weber, E. W. Myers. Human Whole-Genome Shotgun Sequencing. *Genome Research*, vol. 7, p. 401-409, 1997.

A. N. Whitehead et B. Russel. *Principia Mathematica*. 3 volumes, Cambridge University Press, 1910-1913.

A. Wuensche. Genomic Regulation Modeled as a Network with Attraction Basins. *Proceedings of the Pacific Symposium on Biocomputing*, p. 89-102, 1998.

R. Zaidel-Bar, S. Itzkovitz, A. Ma'ayan, R. Iyengar, B. Geiger. Functional Atlas of the Integrin Adhesome. *Nature Cell Biology*, vol. 9, 858-867, 2007.

R. Zaidel-Bar, Z. Kam, B. Geiger. Polarized Downregulation of the Paxillin-p130^{CAS}-Rac1 Pathway Induced by Shear Flow. *Journal of Cell Science*, vol. 118, p. 3997-4007, 2005.

E. Zamir, B. Geiger. Molecular Complexity and Dynamics of Cell-Matrix Adhesions. *Journal of Cell Science*, vol. 114, p. 3583-3590, 2001.

Annexes

{Note à destination de l'éditeur : les annexes présentées ici devraient apparaître comme des encarts dans le corps du texte.}

Annexe 1 : Graphe

{Note à destination de l'éditeur : encart à placer au début de la partie 2.3 traitant de la bio-informatique des réseaux.}

Historiquement, le concept de graphe est introduit par Euler au 18ème siècle au travers du problème des sept ponts de la ville de Königsberg, aujourd'hui rebaptisée Kaliningrad. Ce problème se pose très simplement de la façon suivante : (hypothèse) Königsberg est bâtie autour de deux îles de la Pregolia reliées par un pont et six autres ponts permettent de rejoindre l'une ou l'autre de ces îles depuis les rives de la Pregolia ; (question) pour n'importe quel point de départ dans Königsberg, existe-t-il un itinéraire passant une unique fois par chacun des sept ponts permettant de retourner au point de départ choisi ? Pour résoudre ce problème (Euler, 1741), Euler le modélise sous forme de graphe en représentant les îles et les rives par des points et les ponts par des traits reliant les points les uns aux autres comme le montre la Figure A1.

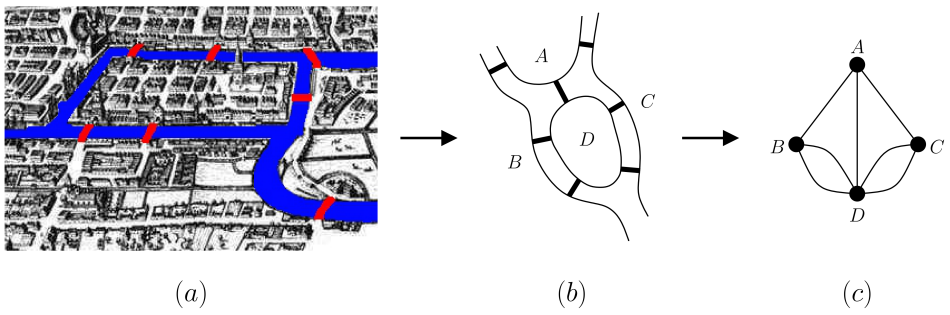

Fig. A1 : *(a)* Représentation de Königsberg, de la Pregolia, de ses deux îles, de ses rives et de ses sept ponts ; *(b)* Simplification schématique en séparant la ville en quatre régions, avec *A* et *D* qui représentent les îles, et *B* et *C* les rives ; *(c)* Modélisation sous forme de graphe.

En théorie des graphes (qu'on peut voir aussi bien comme une branche des mathématiques ou de

l'informatique), les points sont appelés des *sommets* et les traits des *arêtes*. Plus formellement, un graphe G est donc un couple (S, A), où S représente l'ensemble des sommets et A l'ensemble des arêtes. En toute généralité, un graphe peut avoir plusieurs arêtes différentes qui relient la même paire de points. A est alors un multi-ensemble contenant soit des paires de sommets (dans le cas d'un graphe dit *non orienté*), soit des couples de sommets (dans le cas d'un graphe dit orienté). Mathématiquement, rappelons qu'une paire est un ensemble composé de deux éléments alors qu'un couple est un ensemble ordonné (c'est-à-dire un vecteur) composé de deux éléments. Informellement, contrairement au cas d'un graphe non orienté, dans un graphe orienté, les arêtes sont dirigées (et sont d'ailleurs souvent appelées *arcs*). La figure A2 illustre ces concepts.

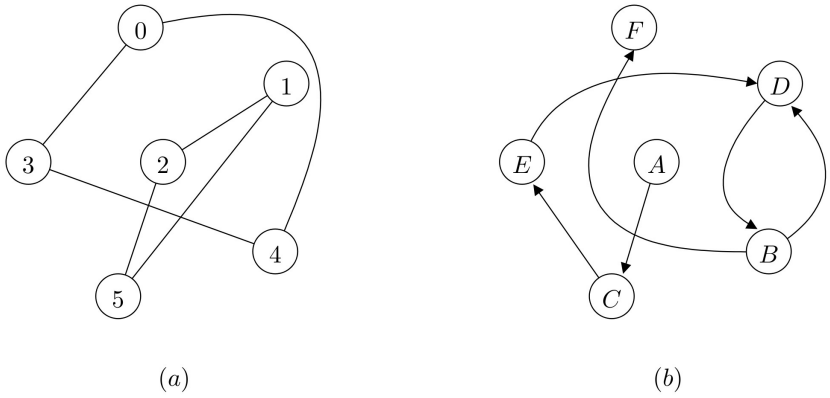

Fig A2 : *(a)* Un graphe non orienté $G = (S, A)$, avec $S = \{0, 1, 2, 3, 4, 5\}$ un ensemble de six sommets et $A = \{\{0, 3\},\{0, 4\}, \{1, 2\}, \{1, 5\}, \{2, 5\}, \{3, 4\}\}$ un ensemble de six arêtes ; *(b)* Un graphe orienté $G' = (S', A')$, avec $S' = \{A, B, C, D, E, F\}$ un ensemble de six sommets et $A' = \{(A, C), (B, D), (B, F), (C, E), (D, B), (E, D)\}$.

Annexe 2 : Arbre

{Note à destination de l'éditeur : encart à placer au début de la partie 2.5 traitant de la phylogénie.}

Un arbre est un graphe connexe acyclique, à savoir un graphe dans lequel chaque sommet est accessible à partir de n'importe quel autre (connexité) et dans lequel il n'existe pas de cycle, c'est-à-dire une succession d'arêtes $(a_1, \ldots, a_{k-1})$, avec $k \geq 2$, telle que $a_{k-1} = a_1$. Formellement, un arbre est donc un graphe $G = (S, A)$, où S représente l'ensemble des sommets et A l'ensemble des arêtes (une arête étant elle-même définie comme une paire de sommet), défini récursivement de la manière suivante :

- un sommet $s \in S$ est un arbre,
- G est un arbre si et seulement si $(S \cup s', A \cup \{\{s', s\}\})$, où $s \in S$ et $s' \notin S$ est un élément quelconque, est un arbre également.

En informatique, les arbres sont des graphes orientés connexes et acycliques. Au même titre que les tableaux (ou vecteurs dans le langage mathématique), ils sont largement utilisés en tant que structure de données. Ils permettent notamment d'avoir une représentation hiérarchisée des données ce qui est souvent particulièrement utile du point de vue algorithmique. À titre d'exemple, la Figure A3 présente sous forme de tableau trié (par ordre croissant) l'ensemble des quinze premiers nombres entiers naturels et son équivalent sous la forme d'un arbre binaire de recherche équilibré de type A.V.L.[9]. Avec ce type d'arbre, l'on sait que la complexité dans le pire cas des opérations de recherche, d'insertion et de suppression d'un élément est logarithmique, contrairement au cas d'un tableau trié, dans lequel il faudra compter sur une complexité linéaire pour les opérations d'insertion et de suppression dans le pire cas (Cormen, Leiserson et al., 1994).

9 Dans un arbre, on appelle feuille un sommet de l'arbre qui ne possède pas de successeur. Par exemple, l'arbre AVL de la Figure 4 admet 8 feuilles : 0, 2, 4, 6, 8, 10, 12 et 14. *A contrario*, on appelle racine le sommet de l'arbre qui ne possède pas de prédécesseur. 7 est par conséquent la racine de l'arbre de type A.V.L. de la Figure 4.

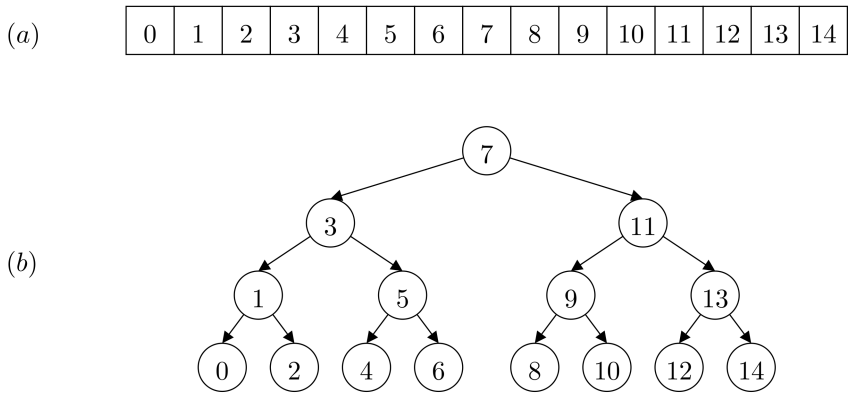

Fig. A3 : *(a)* Tableau trié par ordre croissant contenant les quinze premiers nombres entiers naturels. *(b)* Sa représentation sous forme d'arbre de type A.V.L., c'est-à-dire tel que chacun de ses sommets contient dans son sous-arbre gauche (resp. droit) des valeurs qui sont inférieures (resp. supérieures) à la sienne et que les hauteurs des deux sous-arbres d'un même sommet diffèrent d'au plus un.

Dans les arbres, nous distinguons généralement différents types de sommets, qui sont par ailleurs appelés nœuds la plupart du temps en référence à la métaphore de l'arbre vivant. Ainsi, les nœuds qui sont tels qu'il n'admettent pas de sous-arbre sont appelés les *feuilles*. *A contrario*, celui qui est tel qu'aucun autre nœud ne l'admet comme élément d'un de ses sous-arbres est appelé la *racine* de l'arbre. Dans la Figure 4.(b), les feuilles sont les nœuds 0, 2, 4, 6, 8, 10, 12, 14, la racine, unique, est quant à elle le nœud 7.

Annexe 3 : une application aux résultats étonnants, entre comparaison de séquences et phylogénie

{Note à destination de l'éditeur : encart à placer à la fin de la partie 2.5 traitant de la phylogénie.}

Entre autres résultats significatifs, les études en phylogénie ont mis en évidence le transfert horizontal de gènes, c'est-à-dire le transfert d'une espèce à une autre. Le qualificatif "horizontal" est relatif aux arbres phylogénétiques, dans lesquels les nœuds localisés à une même hauteur correspondent à des espèces contemporaines. La découverte de ce transfert horizontal de gènes explique notamment la résistance aux antibiotiques de certaines bactéries. Cependant, outre cette résistance bactérienne, un exemple de résultats parmi les plus impressionnants est certainement celui de la syncytine, une protéine indispensable au développement de l'embryon, sans laquelle la gestation complète est rendue impossible (Dupressoir, Vernochet et al., 2009). Étonnamment, cette protéine ressemble "trait pour trait" à un gène de rétrovirus, gène qui proviendrait d'infections passées de cellules germinales d'individus d'espèces ancestrales, infections causées par des rétrovirus qui auraient laissé, au sein du génome humain, une partie de leur propre génome. Ainsi, les recherches dans ce domaine ont montré que le gène Env viral, retrouvé à la fois chez l'homme et la souris, code pour une protéine de l'enveloppe virale qui intervient au niveau de la phase d'accrochage des virus à la membrane cellulaire lors de l'infection. Par ailleurs, le rôle connu des protéines d'enveloppe des rétrovirus endogènes dans la placentation appuie l'idée que la présence de ces virus participent à notre état de mammifères. De là à dire que, sans les virus, nous ne serions pas des mammifères, il n'y a qu'un pas… Dans (Dupressoir, Vernochet et al., 2009), les auteurs ont effectivement identifié des propriétés immunosuppressives de domaines très conservés portés par les protéines d'enveloppe de ces rétro-éléments et qui jouent un rôle essentiel dans la capacité des rétrovirus à envahir leur hôte, la capacité des cellules tumorales à échapper à la réponse immune anti-tumorale et dans la formation de la barrière materno-fœtale du placenta des mammifères.

Une partie considérable du génome des vertébrés est composée d'éléments répétés d'origine

rétrovirale, les rétrovirus endogènes, intégrés au génome il y a plusieurs millions d'années lors d'infections de la lignée germinale. Certains d'entre-eux ont conservé un ou plusieurs de leurs gènes intacts. C'est le cas du gène d'enveloppe (Env). Pour le montrer, c'est une recherche systématique dans les génomes séquencés qui a permis d'identifier le gène de la syncytine, acquis de manière indépendante il y a vingt à quarante-cinq millions d'année par les génomes murins et humains, et qui présentent une expression placentaire et peuvent faire fusionner les cellules entre elles. Toutefois, comparer les génomes et les séquences n'a pas suffi à "prouver" (le concept de preuve en biologie interpelle souvent les scientifiques des sciences dites "dures" mais ceci est un sujet qui pourrait faire l'objet d'un chapitre à part entière) l'hypothèse originale de l'équipe de Heidmann. Il a fallu y joindre de la biologie expérimentale, grâce à des souris mutées (forçant l'inhibition de la syncytine), afin de mettre en évidence que l'absence de ce gène empêchait des souris de mener à terme leur gestation (Dupressoir, Heidmann, 2011). Il apparaît ainsi que la capture de rétrovirus endogènes a été un événement majeur dans l'établissement et l'évolution de la structure placentaire.