STATISTIQUES ET GÉNOME

En guise d'introduction

Franck Picard [1]

1. L'introduction de l'introduction

Alors que nous fêtons les dix ans du séquençage du génome humain, c'est avec plaisir que nous nous regroupons pour partager avec la *Gazette des mathématiciens* notre expérience dans le domaine des statistiques appliquées à la génomique. En guise d'introduction, nous avions envie de retracer l'historique des échanges entre ces deux disciplines. C'est un aperçu partiel, biaisé, succinct, qui fera sûrement hurler bon nombre de spécialistes de tel ou tel sujet, mais nous espérons que son fil conducteur aidera les lecteurs à comprendre l'articulation des problèmes statistiques rencontrés sous la lumière des avancements et des grandes orientations de la biologie moléculaire de ces trente dernières années. Plus modestement, j'espère que cette introduction permettra de comprendre les liens entre les différents articles de ce dossier !

2. Comment faisait-on avant l'ADN ?

On pourrait considérer la génomique comme une branche récente de la génétique. En effet, les deux disciplines ont pour objectif central d'élucider les fondements biologiques de l'hérédité, et pour principe commun d'étudier les liens entre variations de caractères observés (ou phénotype), et variations de facteurs héritables (génotype). La distinction des deux branches est avant tout historique. Jusqu'à la découverte de la molécule d'ADN dans les années cinquante, le concept de gène n'était fondé que sur des observations de transmissions de caractères et n'avait pas de réalité physique. Les travaux fondateurs de Mendel et Morgan montrent que les facteurs héritables sont transmis par des entités distinctes qui sont organisées linéairement (mais on ne sait pas sur quel support). Morgan montre aussi que la distance entre les gènes détermine leur transmission d'une génération à une autre. À cette époque, la génétique est plutôt une branche de la statistique, avec notamment les méthodes de cartographie génétique s'appuyant sur les modèles linéaires et sur les modèles probabilistes de génétique des populations dans la lignée des premiers développements de Fisher. Le terme de génome est employé pour la première fois en 1932 pour désigner l'intégralité de l'information génétique d'un organisme. Alors que la génétique est profondément ancrée dans l'étude des données populationnelles, la génétique moléculaire et la génomique

[1] LBBE, UMR CNRS 5558 Université Lyon 1, F-69622, Villeurbanne, France.

s'intéresseront plus particulièrement à la molécule d'ADN et à ses propriétés. La génomique en tant que science connaît un essor considérable à partir des années 90 avec la disponibilité des génomes complets.

3. Seulement 4 lettres ! (oui mais enroulées en double hélice)

La transition de la génétique vers la génomique commence par l'exploration des bases moléculaires de la génétique. C'est grâce à deux étapes fondamentales que l'on commence à matérialiser la notion de gène. Beadle et Tatum montrent d'abord que le gène constitue l'information à la source de molécules individuelles dans une voie de régulation biologique, et posent le postulat « un gène une enzyme » [6]. Ensuite on démontre que l'hérédité possède un support physique d'une nature chimique différente des protéines [12, 5]. C'est grâce à la résolution de la structure 3D [33] de la molécule d'ADN que l'on obtient la clé pour expliquer par quels mécanismes l'ADN peut être la molécule vecteur de l'hérédité. L'ADN – Acide Désoxyribo Nucléique – est une molécule double brin dont chaque brin est un polymère orienté composé de 4 molécules élémentaires (nucléotide), Adénine, Thymine, Cytosine, Guanine, qui sont reliées de manière covalente (forte). Chaque nucléotide a la propriété de s'apparier de manière non covalente (faible) avec un autre nucléotide, et ce de manière spécifique, A-T, G-C, c'est le principe d'appariement des bases qui permet de déduire le deuxième brin du premier. Cette propriété de complémentarité explique à elle seule comment un brin peut servir de modèle pour la synthèse d'un autre brin, ouvrant ainsi la voie aux mécanismes de la réplication, qui est à la source de tout mécanisme de transmission. De plus, l'existence de deux brins complémentaires et séparables permet d'expliquer comment des erreurs de réplication peuvent conduire à des mutations transmises à la descendance par l'intermédiaire de l'une des deux molécules filles. La découverte de la structure de la molécule d'ADN offre donc les bases moléculaires des théories de l'évolution. Elle explique également les liens précoces de la génomique et de l'algorithmique du texte. En 1975 Frederick Sanger annonce qu'il a mis au point une technique permettant de déterminer la succession des nucléotides le long de la molécule d'ADN [22]. C'est le début du séquençage des gènes et de leur modélisation par un texte écrit dans un alphabet de 4 lettres : `A`, `T`, `G`, `C`. Le séquençage des génomes entiers commencera dans les années quatre-vingt-dix.

4. Le développement des bases de données

Dès la fin des années soixante-dix se mettent en place des initiatives nationales et internationales pour stocker et cataloguer systématiquement les informations de séquences afin de les rendre disponibles à la communauté scientifique, à travers des initiatives comme GeneBank et l'EMBL [26, 8]. C'est le début d'un « naturalisme moderne » qui consiste à collecter les génomes comme on a collecté les spécimens aux XVIII^{e} et XIX^{e} siècles [26]. L'explosion de la micro-informatique et la popularisation des micro-ordinateurs ont alimenté l'attrait pour l'analyse automatique des séquences d'ADN, parce que d'une part les séquences sont illisibles sans modèle, et d'autre part les méthodes mathématiques et informatiques étaient de toute évidence la seule manière de faire face au déluge de séquences produites par les projets de séquençage. Les progrès toujours constants des techniques de séquençage posent

désormais d'importants défis informatiques aux banques de données publiques, à tel point que la survie même du modèle de base de données centralisée est remise en question à l'heure actuelle. Alors qu'il a fallu dix ans pour produire une première ébauche du génome humain (3,5 milliards de nucléotides), le projet 1000 Génomes propose par exemple de cataloguer les variations génétiques de 1000 individus sur la totalité de leur génome. Aujourd'hui, séquencer un génome humain peut se faire en quelques mois par un seul laboratoire pour quelques milliers d'euros. En avril 2011, GenBank contenait 126 551 501 141 nucléotides dans 135 440 924 séquences. De nouvelles initiatives voient donc le jour pour faire face à des jeux de données qui atteignent désormais le peta-byte (10^{15}) [8].

5. Les débuts de la bioinformatique

Le concept de bioinformatique émerge dans les années soixante-dix, avec l'idée sous-jacente que les sciences de l'information pouvaient aider à la compréhension des systèmes biologiques. Un axe historique des développements mathématiques pour l'étude du génome est la génomique comparative. L'idée centrale est que les séquences biologiques (ADN et protéines) contiennent des traces de l'évolution qui sont étudiées par comparaison. L'hypothèse de base est que les dissimilarités entre séquences proviennent de mécanismes évolutifs comme les insertions de nucléotides, les délétions, et les substitutions. Se développent alors des modèles mathématiques d'évolution moléculaire [14, 16] permettant la reconstruction d'une histoire évolutive en utilisant les processus stochastiques et la combinatoire des arbres phylogénétiques.

Le calcul de similarité entre séquences s'effectue par des techniques d'alignement pour déterminer si deux chaînes de caractères se ressemblent. Ce sont les travaux (entre autres) de Needleman et Wunsch [19], Sankoff [22], et Smith et Waterman [32] dans les années soixante-dix qui donnent des solutions algorithmiques à ce problème d'alignement de séquences, à l'aide de la programmation dynamique qui permet de calculer des scores optimaux d'alignements. Le score le plus simple serait par exemple de noter 1 si deux lettres correspondent et 0 sinon, puis ensuite de chercher les segments de séquences de score élevé. En 1990 est publié BLAST (Basic Local Alignment Search Tool, [3]), qui propose à la communauté un outil de recherche de séquences dans les banques. C'est une des publications les plus citées dans la littérature scientifique des années quatre-vingt-dix. Mais très vite se pose une question cruciale : quel sens donner au score d'alignement de deux séquences ? Ce score est-il significatif ? Le point de vue statistique mettra l'accent sur l'aspect stochastique de l'évolution des séquences en proposant un formalisme probabiliste aux méthodes d'alignements avec notamment les travaux de Samuel Karlin et Volker Brendel [15]. Le principe de l'approche statistique de l'alignement est de tester si le score observé est plus élevé qu'attendu sous un modèle aléatoire. Les modèles de Markov permettent alors de définir un modèle de référence pour les séquences biologiques, et donnent alors accès à la loi du score d'alignement. La notion d'exceptionnalité du score est quantifiée en calculant la probabilité que le long d'une séquence, le score du segment de meilleur alignement dépasse un seuil. Ces développements font appel à la théorie des maxima de sommes partielles de processus. La notion d'exceptionnalité s'applique également

à la sur/sous-représentation de certains mots dans les séquences [21], avec l'idée sous-jacente qu'une chaîne de caractères (motif) exceptionnellement présente dans un texte pourrait avoir un sens biologique particulier. Ces développements feront l'objet de l'article de S. Schbath.

Les premières contributions des statistiques à la génomique s'inscrivent également dans le courant de l'analyse automatique du langage, qui se développe dans les années soixante-dix, avec la mise au point et l'utilisation des modèles de chaînes de Markov cachées qui permettent de modéliser et de localiser des changements dans les fréquences d'apparition des nucléotides le long de la séquence. Ces modèles probabilistes permettent alors l'annotation automatique des génomes qui constitue une dynamique majeure des années quatre-vingts. Cet effort considérable de cartographie physique permet de découvrir, de positionner les gènes le long de l'ADN et de les munir d'une signature de séquence. Il contribue à la vision des années soixante/soixante-dix du génome : un code présent dans l'ADN qui donne lieu à des produits fonctionnels.

6. Après la séquence ?

Alors que les projets de séquençage permettent d'établir des modèles de fonctionnement et d'évolution des génomes de plus en plus précis, la notion de gène se complexifie de plus en plus. Un gène devient un segment d'ADN qui contribue au phénotype ou à une fonction biologique. Les techniques d'annotations montrent que la structure des gènes est extrêmement complexe. Les premières structures élucidées étaient les plus simples, mais il apparaît rapidement que les gènes ont des structures éclatées en plusieurs morceaux (plusieurs exons). Si un gène est constitué de trois exons A, B, C, l'agencement de ces exons peut produire des transcrits ayant des fonctions différentes. Ce phénomène s'appelle l'épissage alternatif : alors que la quantité de gènes annotés n'augmente quasiment plus aujourd'hui chez l'homme, c'est le nombre de formes différentes par gène qui explose. Aujourd'hui les annotations contenues dans GENCODE [13] contiennent une moyenne de 5, 4 formes différentes par gène (transcrits alternatifs), et plus de la moitié des gènes a un site d'initiation de la transcription alternatif. Alors que le texte contenu dans la séquence d'ADN est déterminé, les produits de l'expression de ce texte apparaissent extrêmement divers. Par conséquent, une des conclusions majeures des projets de séquençage est que l'information de séquence ne constitue qu'une part (certes fondatrice) de la complexité des bases moléculaires des phénomènes biologiques étudiés. Alors que dans les années 2000 les génomes de plus de 800 organismes ont été séquencés, le déséquilibre est criant entre le nombre de séquences stockées et la connaissance que l'on a de leurs fonctions biologiques. L'élan qui succède aux projets de séquençage a pour objectif de s'intéresser à grande échelle, en plus du génome, aux produits provenant de l'expression des séquences génomiques : les ARN et les protéines.

Le dogme central de la biologie moléculaire est posé par Francis Crick dans les années soixante [27] et propose un modèle de passage de l'information de la séquence d'ADN aux protéines par les ARN (Acide Ribo Nucléique). Dans ce modèle, l'ADN constitue la matrice qui contient l'information qui est transcrite dans

un autre alphabet de 4 lettres (la Thymine -T- est remplacée par l'Uracile -U) dont le support sont les « transcrits » composés d'ARN. Les protéines constituent les molécules « exécutantes » des fonctions biologiques, et l'ARN la molécule de transition (le messager) entre les acides nucléiques et les protéines. Ce sont d'ailleurs les ARN messagers qui sont « traduits » en protéines à l'aide du code génétique. Après avoir étudié le « texte » que constituent les molécules d'ADN, une extension naturelle de la génomique était l'étude des produits des gènes. Pour des raisons techniques, il est plus facile d'étudier les ARN messagers que les protéines, et les techniques classiques ne permettaient d'étudier la présence que de quelques dizaines d'ARN différents simultanément. C'est l'émergence de la technologie des puces à ADN dans les années quatre-vingt-dix qui permet l'essor de la biologie dite « à haut débit ».

7. La biologie à haut débit et « petits » problèmes statistiques associés

La technologie des puces à ADN (DNA microarrays) repose sur la propriété de complémentarité des bases de la molécule d'ADN. Considérons la portion d'ADN correspondant à la séquence d'un gène. Cette molécule comporte deux brins complémentaires pouvant être séparés. On appelle hybridation spécifique la capacité d'un simple brin à « retrouver » son brin complémentaire. Simplifiée à l'extrême, la technologie des puces à ADN consiste à fixer un catalogue de séquences connues simple brin sur une lame de verre, et à déposer sur cette lame un extrait des ARN d'une cellule (pour des raisons techniques ces ARN sont convertis en ADN). Si sur la lame sont déposées les séquences correspondant aux transcrits des gènes A,B,C, et que les produits des gènes A,C,D,E sont présents dans l'échantillon, alors la puce à ADN permettra de quantifier la présence des produits des gènes A et C, les produits des gènes D et E n'étant pas détectés sur la lame (car non prévus dans le catalogue). Historiquement, la technologie des puces à ADN a été mise au point grâce aux transferts de technologie provenant des circuits imprimés des cartes à puce [18], et les projets de séquençage ont permis de créer des catalogues extrêmement larges. Les deux technologies qui s'imposent alors sont mises au point par Pat Brown à Stanford [24], et par la société Affymetrix [10]. Elles reposent sur la fixation sur des supports fixes de milliers de séquences connues, et sur la mesure de la quantité de séquences hybridées sur la lame par des signaux de fluorescence.

C'est l'essor du domaine de la biologie à haut débit et des sous-disciplines en « -omique » : on caractérise les cellules non seulement grâce à leur séquence d'ADN (génomique), mais aussi grâce à l'ensemble de leurs transcrits et de leurs protéines (transcriptomique, protéomique). La contribution des statistiques à ce domaine est alors centrale. En effet la technologie des microarrays produit des signaux de fluorescence (continus) dont l'analyse relève naturellement des statistiques : normalisation des données, analyse des différences d'expression entre conditions (tests), planification expérimentale, classification des gènes et/ou des individus, prédiction et apprentissage statistique. On considère désormais l'ensemble des produits d'expression des gènes d'un organisme pour identifier des signatures moléculaires. Les deux études emblématiques du début de la transcriptomique concernent la prédiction de deux types de leucémies sur la base de l'expression des gènes des patients [11], et

la distinction de sous-types de cancers sur des critères moléculaires [2], ouvrant la voie au diagnostic moléculaire des maladies, et en particulier des cancers. La technologie des microarrays est aussi étendue à d'autres problématiques, comme la cartographie des anomalies chromosomiques ou des interactions entre l'ADN et certaines protéines.

C'est l'époque du lancement du projet ENCODE [7] qui, à l'image des grands projets de séquençage, a pour objectif de fournir une encyclopédie du fonctionnement du génome humain (ENCyclopedia Of DNA Elements). Cette époque voit également se populariser l'utilisation de méthodes statistiques jusqu'à présent absentes des laboratoires de biologie moléculaire. C'est même l'ère de l'utilisation quasi-obligatoire des ordinateurs et des logiciels d'analyse (autres qu'Excel !) afin de pouvoir analyser des données comportant plusieurs dizaines de milliers d'enregistrements. Du côté des statisticiens, c'est le développement considérable du logiciel libre R[2] qui, grâce à l'essor d'internet permet aux statisticiens de diffuser leurs méthodes sous forme de logiciels (packages).

Cependant, même si les tâches statistiques sont bien identifiées, les techniques classiques reposent sur le postulat que le nombre d'observations n était beaucoup plus grand que le nombre de variables p. Cette tendance s'inverse dans les années 1990-2000 car il devient beaucoup moins coûteux de mesurer plusieurs milliers de variables sur peu d'individus par les techniques à haut débit, que de trouver un échantillon représentatif sur lequel fonder des prédictions solides. Les fondements asymptotiques des statistiques classiques ne sont donc plus valides, et les tâches autrefois simples deviennent un enjeu méthodologique majeur. Les microarrays ne sont qu'un exemple (avec le nombre de gènes étudiés correspondant au nombre de variables) et l'analyse d'images ou de courbes rencontre le même problème. Ces problématiques statistiques s'inscrivent dans une dynamique très populaire dans les années 1990-2000, celle de l'apprentissage statistique, qui regroupe des chercheurs en informatique, mathématique et statistique sous l'impulsion (entre autres) des travaux de Vapnik et Chervonenkis dans les années soixante-dix [29]. L'objectif est de construire un classifieur (ou règle de décision) prenant en compte les descripteurs pour prédire une étiquette d'appartenance à une classe ou label. La construction du classifieur se fait sur un échantillon pour lequel les labels sont connus (échantillon d'apprentissage). En génomique, on cherchera à prédire le statut biologique d'un patient (son label) à l'aide de ses caractéristiques génomiques (une variable quantifie la quantité d'expression d'un gène). Un élément crucial de la construction d'un classifieur performant est bien entendu la sélection des variables les plus informatives/prédictives. Le thème de la sélection de variables est également un domaine extrêmement actif de ces dernières années, avec la mise au point de techniques dites de « régression creuse » (sparse regression) sous l'impulsion des travaux de R. Tibshirani [28]. Au vu de l'importance des développements méthodologiques suscités par l'essor de l'étude des données de puces à ADN, les articles de E. Lebarbier et P. Neuvial de ce dossier y seront consacrés.

[2] `http://www.r-project.org/`

8. Une médecine personnalisée ?

On assiste alors au développement de la « translational research » qui consiste à transférer le plus rapidement possible les avancées de la recherche fondamentale à la recherche clinique pour une meilleure prise en charge du patient. Alors que des budgets considérables sont mis en œuvre (notamment pour la recherche contre le cancer), et après une succession d'annonces sur les promesses de la médecine moléculaire personnalisée et sur l'identification « de biomarqueurs » (« Array of Hope » [17]), les statisticiens ont souvent eu un rôle de modérateur, voire de rabat-joie. En effet, malgré les progrès des méthodes statistiques pour la prédiction, si l'échantillon d'apprentissage est trop petit, les règles de prédictions montreront des performances modestes sur de nouveaux individus. De plus les études moléculaires ont mis en lumière une incroyable variabilité inter-individuelle concernant les maladies complexes. À ce titre, les années 2000 ont vu progresser le nombre de publications rapportant des résultats non reproductibles, avec des titres évocateurs comme « The cancer biomarker problem » [23], ou « Valid Concerns » [4] ! Malgré ces limitations, la technologie des microarrays a cependant eu un impact considérable dans l'appréhension des phénomènes moléculaires à l'échelle du génome entier avec des protocoles utilisables par chaque laboratoire. La technologie des microarrays est maintenant utilisée en routine et permet d'étudier des phénomènes moléculaires sans *a priori* sur le gène d'intérêt. Au travers d'une utilisation plus rigoureuse des statistiques par l'ensemble de la communauté des « génomiciens » nous avons également pu faire passer l'idée qu'il était crucial de distinguer la variabilité technique (qui peut être contrôlée par des procédés expérimentaux plus fiables) de la variabilité biologique, dont la quantification requiert un nombre minimum d'individus. Le sur-optimisme des études prédictives est sans aucun doute lié à la séparation des communautés de la génétique quantitative et de la génomique, les uns ayant peu l'habitude des données à grande dimension, les autres ne connaissant pas forcément les principes fondamentaux des études génétiques. On voit maintenant émerger le concept de génomique-génétique ou de génomique des populations : le retour des problématiques (classiques ?) de génétique étudiées avec les outils et au regard des connaissances apportées par la génomique. On peut désormais superposer différentes complexités, moléculaires et populationnelles.

9. Après les transcrits ? C'est complexe !

De même que la progression des projets de séquençage a mis en lumière le besoin de comprendre le monde des ARN, après dix années de projets « post-génomiques », un changement de mode de pensée était nécessaire pour appréhender l'ensemble des données collectées. Alors que les grands projets de génomique cataloguent les bases moléculaires des génomes et de leur expression, progressivement les nouveaux projets ont adopté un nouvel angle d'attaque en se focalisant directement sur les fonctions biologiques dans leur ensemble. Or ces fonctions sont généralement le résultat d'interactions entre protéines, entre voies de signalisation et de régulation. Plutôt que d'envisager les strates d'information une par une, l'enjeu réside désormais dans la compréhension des interactions entre ces différentes « échelles de complexité ».

La complexité ! Les réseaux ! Ce sont certainement les deux mots clés de ces dernières années. Selon l'ISI Web of Knowledge le champ thématique des « complex networks » était la principale thématique de recherche en mathématiques en 2008. Cette dynamique provient d'un élan commun à plusieurs disciplines : l'explosion des réseaux sociaux et l'intérêt grandissant pour l'étude de leurs structures, la disponibilité des données en physique des particules, et la quantité considérable d'informations collectées en biologie moléculaire. La distinction individus/variables est désormais trop restrictive : les données sont constituées d'agents dont les interactions permettent le fonctionnement d'un système. Les fondements de la biologie des systèmes sont donc ancrés dans l'intégration des informations récoltées par les projets de génomique, afin de modéliser les processus biologiques, et de concevoir aussi de nouvelles expériences. On assiste à une implication croissante de la physique dans cette discipline, qui voit la cellule comme un système contrôlable, et qui ouvre la voie à une véritable biologie in-silico. Les graphes constituent l'objet mathématique central de l'analyse des réseaux biologiques, ce qui explique l'implication extrêmement forte des mathématiques discrètes. La statistique quant à elle permet l'inférence des mécanismes de régulation (graphes de régulation), dans un contexte de ultra-haute dimension (la taille du problème explose potentiellement avec le nombre de gènes [30]). Les phénomènes de régulation sont aussi modélisés par des réseaux booléens en grande dimension. Les statisticiens se sont également intéressés à la détection de structures particulières dans les réseaux biologiques. On parle de « modules » par exemple, qui regroupent des gènes présentant un fonctionnement homogène du point de vue d'une fonction donnée. On peut également rechercher des structures de connectivité au sens large [9, 1]. La notion de motif, axe phare de l'analyse des séquences, a également été généralisée aux réseaux biologiques [25, 20], avec l'idée sous-jacente que certaines sous-structures pouvaient être à l'origine du fonctionnement de l'ensemble du réseau. Le thème de l'analyse statistique des réseaux biologiques fera l'objet du dernier article de ce dossier.

10. Et après ?

Décidément, nous assistons à une révolution tous les 5 à 10 ans dans le domaine de la génomique ! Aujourd'hui, c'est le grand retour des technologies de séquençage qui crée un nouvel engouement. Et pour cause : après les microarrays, la nouvelle transition technologique permet de séquencer des génomes d'organismes entiers en quelques semaines pour « quelques » milliers d'euros ! Des consortiums internationaux nous ont déjà permis d'accéder aux génomes de $\sim$250 eucaryotes (organisme possédant un noyau cellulaire), $\sim$4000 bactéries et virus, avec des génomes complets pour l'homme, la souris, le rat, le chien, le chimpanzé, la vache, mais aussi pour un marsupial, l'ornithorynque et pour un oiseau ! La nouvelle révolution est que ce séquençage de génomes complets est maintenant réalisable par des laboratoires de taille modeste, et pas uniquement pour des organismes modèles. Cet accès grandissant aux technologies à haut (très haut) débit permet de sortir du modèle de « génome de référence » : les projets internationaux cherchent désormais à étudier la diversité génétique à l'échelle du génome entier et de populations. C'est le cas du projet 1000 génomes par exemple [31].

En tant que statisticiens, nous savons d'expérience qu'il est très difficile de prédire les grandes tendances des prochaines années ! Sans trop prendre de risques,

on peut néanmoins supposer que la thématique de la « grande dimension » sera à l'honneur, avec des quantités de données à modéliser qui rivalisent avec les quantités de la physique des particules. Peut être un rapprochement entre l'informatique, les mathématiques et la statistique ? L'accès, le stockage, la manipulation est déjà problématique, et les statisticiens sont souvent moins bien armés que d'autres pour modéliser des données à structures exotiques (par rapport à la théorie des graphes par exemple). Une bonne occasion pour encourager nos collègues mathématiciens qui souhaiteraient nous rejoindre dans l'aventure ! Et aussi pour vous souhaiter bonne lecture de ce dossier !

11. Références

[1] E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing. Mixed Membership Stochastic Blockmodels. *J Mach Learn Res*, 9 :1981–2014, Sep 2008.

[2] A. Alizadeh and al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 403 :503–511, Feb 2000.

[3] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *J. Mol. Biol.*, 215 :403–410, Oct 1990.

[4] No authors listed. Valid concerns. *Nature*, 463 :401–402, Jan 2010.

[5] O. T. Avery, C. M. Macleod, and M. McCarty. Studies on the chemical nature of the substance inducing transformation of pneumococcal types : induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III. *J. Exp. Med.*, 79 :137–158, Feb 1944.

[6] G. W. Beadle and E. L. Tatum. Genetic Control of Biochemical Reactions in Neurospora. *Proc. Natl. Acad. Sci. U.S.A.*, 27 :499–506, Nov 1941.

[7] E. Birney and al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 447 :799–816, Jun 2007.

[8] G. Cochrane and al. Petabyte-scale innovations at the European Nucleotide Archive. *Nucleic Acids Res.*, 37 :19–25, Jan 2009.

[9] J.J. Daudin, F. Picard, and S. Robin. A mixture model for random graph. *Statistics and computing*, 18(2) :1–36, 2008.

[10] S. P. Fodor, J. L. Read, M. C. Pirrung, L. Stryer, A. T. Lu, and D. Solas. Light-directed, spatially addressable parallel chemical synthesis. *Science*, 251 :767–773, Feb 1991.

[11] T. R. Golub and al. Molecular classification of cancer : class discovery and class prediction by gene expression monitoring. *Science*, 286 :531–537, Oct 1999.

[12] F. Griffith. The Significance of Pneumococcal Types. *J Hyg (Lond)*, 27 :113–159, Jan 1928.

[13] J. Harrow, F. Denoeud, A. Frankish, A. Reymond, C. K. Chen, J. Chrast, J. Lagarde, J. G. Gilbert, R. Storey, D. Swarbreck, C. Rossier, C. Ucla, T. Hubbard, S. E. Antonarakis, and R. Guigo. GENCODE : producing a reference annotation for ENCODE. *Genome Biol.*, 7 Suppl 1 :1–9, 2006.

[14] T.H. Jukes and C.R. Cantor. *Evolution of protein molecules.* Mammalian protein metabolism. New York : Academic Press., 1969.

[15] S. Karlin and V. Brendel. Chance and statistical significance in protein and DNA sequence analysis. *Science*, 257 :39–49, Jul 1992.

[16] M. Kimura. *The neutral theory of molecular evolution*. Cambridge University Press, 1983.

[17] E. S. Lander. Array of hope. *Nat. Genet.*, 21 :3–4, Jan 1999.

[18] T. Lenoir and E. Giannella. The emergence and diffusion of DNA microarray technology. *J Biomed Discov Collab*, 1 :11, 2006.

[19] S. B. Needleman and C. D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, 48 :443–453, Mar 1970.

[20] F. Picard, J. J. Daudin, M. Koskas, S. Schbath, and S. Robin. Assessing the exceptionality of network motifs. *J. Comput. Biol.*, 15 :1–20, 2008.

[21] G. Reinert, S. Schbath, and M. S. Waterman. Probabilistic and statistical properties of words : an overview. *J. Comput. Biol.*, 7 :1–46, 2000.
[22] F. Sanger. The Croonian Lecture, 1975. Nucleotide sequences in DNA. *Proc. R. Soc. Lond., B, Biol. Sci.*, 191 :317–333, Dec 1975.
[23] C. L. Sawyers. The cancer biomarker problem. *Nature*, 452 :548–552, Apr 2008.
[24] M. Schena, D. Shalon, R. W. Davis, and P. O. Brown. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270 :467–470, Oct 1995.
[25] S. S. Shen-Orr, R. Milo, S. Mangan, and U. Alon. Network motifs in the transcriptional regulation network of Escherichia coli. *Nat. Genet.*, 31 :64–68, May 2002.
[26] B. J. Strasser. Genetics. GenBank–Natural history in the 21st Century ? *Science*, 322 :537–538, Oct 2008.
[27] B. J. Strasser, L. Pauling, and F. Crick. A world in one dimension : Linus Pauling, Francis Crick and the central dogma of molecular biology. *Hist Philos Life Sci*, 28 :491–512, 2006.
[28] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Royal. Statist. Soc B.*, 85(1) :267–288, 1996.
[29] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 2000.
[30] N. Verzelen. Minimax risks for sparse regressions : Ultra-high-dimensional phenomenons. Technical report, Arxiv, 2010.
[31] M. Via, C. Gignoux, and E. G. Burchard. The 1000 Genomes Project : new opportunities for research and social challenges. *Genome Med*, 2 :3, 2010.
[32] M.S. Waterman, T.F. Smith, and W.A. Beyer. Some biological sequence metrics. *Adv. Math.*, 20 :367–387, 1976.
[33] J.D. Watson and F.H.C. Crick. A structure of deoxyribonucleic acid. *Nature*, 171 :964–967, 1953.

Statistiques de motifs

Sophie Schbath [1]

Cet article traite de l'identification de motifs d'ADN fonctionnels le long des génomes par des approches statistiques. Par motif d'ADN, on entend une courte suite de lettres (généralement pas plus d'une quinzaine) dans l'alphabet des nucléotides $\mathcal{A} = \{\mathtt{a}, \mathtt{c}, \mathtt{g}, \mathtt{t}\}$, et par motif fonctionnel, on entend un motif dont les occurrences sur le génome seront reconnues par une protéine qui se fixera alors sur l'ADN pour entrer en action.

L'identification de motifs d'ADN fonctionnels reste un problème biologique encore loin d'être résolu pour plusieurs raisons : (i) leur longueur est variable selon la nature de la protéine, (ii) la reconnaissance de chacune des lettres par la protéine peut être imprécise, autrement dit il n'y a pas nécessairement unicité d'une lettre à chaque position, (iii) l'activité protéique peut dépendre de la présence à proximité d'autres protéines elles-mêmes reconnaissant d'autres motifs ADN, (iv) les motifs fonctionnels ne sont généralement pas conservés d'un organisme à l'autre.

Bien souvent, ces motifs fonctionnels (appelés simplement motifs par la suite) se caractérisent par des répartitions bien particulières le long du génome, d'où la construction de critères ou scores dont on cherchera à mesurer la significativité : on cherche en effet à repérer des événements (ici des occurrences de motifs) qui

[1] INRA, Unité Mathématique, Informatique et Génome, Jouy-en-Josas, France.

auraient très peu de chance de se produire au hasard. Le « hasard » sera déterminé par des séquences aléatoires $X_1X_2\cdots X_\ell$ dont les lettres X_i seront tirées dans l'alphabet $\mathcal{A}$ selon un modèle probabiliste plus ou moins sophistiqué.
Typiquement, on utilise des modèles de chaîne de Markov d'ordre m : la probabilité de générer la lettre $b \in \mathcal{A}$ à la position i dépend des m lettres[2] $a_1a_2\cdots a_m$ qui précèdent, c'est-à-dire aux positions $i-m,\ldots,i-1$. Ces probabilités, dites *de transition*, des lettres $a_1a_2\cdots a_m$ vers la lettre b sont estimées à partir de la composition en mots de taille $m+1$ de la séquence d'ADN analysée. Ce modèle a l'énorme avantage de comparer ce que l'on observe dans la séquence d'ADN étudiée avec ce que l'on pourrait attendre dans des séquences aléatoires ayant en moyenne la même composition en lettres mais aussi en mots de tailles 2, 3, ..., $(m+1)$.

Par exemple, certains motifs se caractérisent par une fréquence anormalement élevée[3] sur le génome entier, ou seulement sur une partie du génome. Pour découvrir d'autres motifs potentiels ayant la même propriété statistique, on est amené à étudier la loi de probabilité du comptage d'un mot dans une chaîne de Markov (cf. section 1) et regarder ceux ayant un comptage significativement élevé (probabilité critique proche de zéro).
D'autres motifs se caractérisent par leur présence dans certaines régions caractéristiques du génome, par exemple en amont des gènes[4]. Si l'on considère alors toutes[5] les sous-séquences de quelques centaines de lettres situées en amont des gènes, cela se traduit par un nombre anormalement élevé de ces sous-séquences contenant au moins une occurrence du motif en question. Dans ce cas, on est amené à évaluer la probabilité qu'un motif donné soit présent (peu importe son nombre d'occurrences) dans une chaîne de Markov (cf. section 2).

D'autres types de questions statistiques relatives aux occurrences de motifs existent mais ne seront pas traitées dans ce dossier. On pourra cependant noter que la formalisation d'un certain nombre d'entre elles utilise non plus un modèle de séquences aléatoires, mais des processus ponctuels pour modéliser les occurrences elles-mêmes. On citera par exemple l'utilisation de processus de Poisson, pour détecter des régions anormalement riches ou pauvres en certains motifs ou pour tester si deux séquences sont aussi riches en un motif donné. Ou encore l'utilisation de processus de Hawkes pour détecter si les occurrences d'un ou plusieurs motifs présentent des distances favorisées ou évitées.

1. Comptage attendu ou anormal ?

Le problème est le suivant : on observe par exemple 762 occurrences du motif de longueur 8 gctggtgg dans une séquence d'ADN de longueur $\ell = 4\,638\,858$ (en fait le génome complet de *E. coli*) et on se demande si ce comptage ne serait

[2] La valeur de m est choisie par le modélisateur et permet de fixer en moyenne la composition des séquences aléatoires jusqu'aux mots de taille de $m+1$.

[3] C'est le cas du motif gctggtgg avec 762 occurrences le long du génome de la bactérie *Escherichia coli* long de $4.6\,10^6$ lettres, ou du motif aagtgcggt avec 740 occurrences le long du génome de la bactétrie *Haemophilus influenzae* de longueur $1.8\,10^6$.

[4] C'est le cas des sites de fixation des facteurs de transcription indispensables à la transcription des gènes en ARN.

[5] Plusieurs milliers, autant que le nombre de gènes par organisme.

pas significativement élevé. Intuitivement, en effet, si les 4^8 mots possibles de taille 8 avaient la même fréquence dans la séquence, on s'attendrait à les observer chacun 72 fois. Pour savoir si l'écart entre 762 (l'observé) et 72 (l'attendu) est significatif, il faut calculer la probabilité de l'événement $\{N \geqslant 762\}$ sous le modèle choisi, appelée *probabilité critique* (ou *p-value* en anglais). On se placera ici dans le modèle de chaîne de Markov d'ordre m ($0 \leqslant m \leqslant h-2$, où h est la longueur du mot étudié) qui permet de s'ajuster sur la composition de la séquence d'ADN en mots de taille 1 à $m+1$.

Le calcul de cette probabilité serait trivial si on connaissait la distribution du comptage N, mais cette dernière est complexe à obtenir du fait que l'on compte des occurrences qui potentiellement peuvent se chevaucher[6] dans la séquence. En effet, le comptage N est une somme de variables aléatoires de Bernoulli Y_i ($Y_i = 1$ si le motif est présent à la position i, et 0 sinon) non indépendantes. Sa loi n'est donc pas une loi binomiale comme l'on aurait pu être tenté de dire.

Plusieurs approches ont été proposées soit pour calculer exactement la probabilité critique soit pour l'approcher. L'une des approches exactes consiste à calculer la distribution du temps d'attente T_n de la n-ième occurrence du mot, pour tout $1 \leqslant n \leqslant 762$, puis d'utiliser le principe de dualité suivant : $\mathbb{P}(T_{762} \leqslant \ell) = \mathbb{P}(N \geqslant 762)$. La distribution exacte du temps d'attente s'obtient par récurrence [4] ou via sa fonction génératrice [10] qu'il convient ensuite de développer en série de Taylor. Néanmoins, la valeur exacte de la probabilité critique n'est réellement calculable numériquement que pour des séquences de quelques dizaines de milliers de lettres et pour des ordres de modèles très faibles, 0 ou 1, ce qui en limite grandement l'usage pour l'analyse de génomes entiers.

Des approximations de la probabilité critique sont donc plutôt utilisées en pratique. Deux types d'approches ont été poursuivies : d'une part celles visant à approcher la distribution du comptage par des lois paramétriques explicites ; d'autre part celles qui approchent directement la queue de la distribution par des approches de grandes déviations [1]. Ces dernières sont en effet pertinentes lorsque les motifs sont très exceptionnels. Parmi les distributions approchées, on peut noter la loi gaussienne pour les motifs plutôt fréquents [2], des lois de Poisson composées pour les motifs plutôt rares [8], voire la loi binomiale pour des motifs non périodiques (dont les occurrences ne peuvent jamais se chevaucher).

Si l'on reprend le problème mentionné au départ de ce paragraphe, il s'avère que la probabilité critique d'observer au moins 762 occurrences du motif `gctggtgg` dans une chaîne de Markov de longueur $\ell = 4\,638\,858$ et ayant en moyenne la même composition en mots de taille 1 à 7 que le génome de *E. coli*, vaut environ $8.7\,10^{-27}$ (en utilisant l'approximation gaussienne). Autrement dit, ce motif est significativement fréquent le long de ce génome (seul deux autres motifs de taille 8 ont une probabilité critique encore plus petite). D'un point de vue biologique, ce motif est vital à la bactérie puisqu'il intervient dans la réparation du génome en cas de cassure des brins d'ADN.

[6] Par exemple, le motif périodique `atgatga` a deux occurrences chevauchantes dans la séquence `gtatgatgatgattc` : l'une à la position 3 (soulignée), l'autre à la position 6 (en italique).

2. Présence attendue ou anormale ?

Ici ce qui nous intéresse est de savoir si le fait qu'un motif soit présent dans une séquence (plutôt courte) soit exceptionnel, c'est-à-dire non dû au hasard, ou pas. Pour cela, on peut bien sûr calculer ou approcher la probabilité $\mathbb{P}(N \geqslant 1) = \mathbb{P}(T_1 \leqslant \ell)$ en utilisant les approches décrites dans le paragraphe précédent.

Dans le cadre de la recherche de sites de fixation de facteurs de transcription, on est très vite amené à considérer l'occurrence de 2 motifs $\mathbf{m}_1$ et $\mathbf{m}_2$ (voire plus) à une certaine distance d plus ou moins variable dans une séquence ($d_1 \leqslant d \leqslant d_2$). Par exemple, `ttgactt` suivi de `ataataa` 16 à 18 lettres après. On s'intéresse donc à l'occurrence du motif composite résultant, noté $\mathbf{m}_1(d_1 : d_2)\mathbf{m}_2$. On pourrait considérer ce motif composite comme un ensemble de longs motifs « simples » (de taille 30 à 32 dans l'exemple) et leur appliquer les résultats précédents, mais la distance est généralement trop grande, produisant un ensemble de motifs « simples » beaucoup trop grand (de l'ordre de 10^{18} dans l'exemple). Il faut donc recourir à des méthodes ad-hoc pour traiter ces motifs composites.

Voici trois approches qui ont été proposées pour évaluer la probabilité qu'un motif composite donné $\mathbf{m}_1(d_1 : d_2)\mathbf{m}_2$ soit présent dans une chaîne de Markov.

– On peut approcher la probabilité $\mathbb{P}(N = 0) = 1 - \mathbb{P}(N \geqslant 1)$ par le produit $(\mathbb{P}(Y_i = 0))^{\ell-h+1}$ où Y_i vaut 1 si le motif composite (i.e. la première lettre du motif) apparaît à la position i dans la séquence, ou 0 sinon, et h est la taille du motif composite [13] ce qui revient à faire comme si les variables Y_i étaient indépendantes. On peut sensiblement améliorer l'approximation en considérant une dépendance d'ordre 1 entre les variables Y_i ce qui ramène à calculer le produit $\mathbb{P}(Y_1 = 0)(\mathbb{P}(Y_i = 0 \mid Y_{i-1} = 0))^{\ell-h}$ [6][7]. Le calcul de $\mathbb{P}(Y_i = 0)$ pour un motif composite est moins trivial que pour un motif « simple » mais s'obtient à partir de la distribution du temps d'attente entre deux motifs « simples » (problème lié au paragraphe précédent). En effet, le motif composite $\mathbf{m}_1(d_1 : d_2)\mathbf{m}_2$ apparaît en position i si et seulement si le premier motif $\mathbf{m}_1$ apparaît en position i et le 2$^{\text{e}}$ motif $\mathbf{m}_2$ apparaît à une distance comprise entre d_1 et d_2 après le 1$^{\text{er}}$ motif. Le calcul de $\mathbb{P}(Y_i = 0 \mid Y_{i-1} = 0)$ est plus technique mais faisable pour un motif composite composé de 2 motifs simples (le travail reste à faire pour 3 motifs ou plus).

– On peut décomposer le temps d'attente avant la première occurrence du motif composite comme une somme aléatoire de temps d'attente indépendants entre motifs simples [11]. Il faut néanmoins faire l'hypothèse que chacun des motifs simples ne peut apparaître plus d'une fois dans le motif composite. Dans le cas du motif composite $\mathbf{m}_1(d_1 : d_2)\mathbf{m}_2$, on obtient la décomposition suivante (cf. figure 1) : il faut attendre le premier motif $\mathbf{m}_1$ puis (**) attendre le prochain motif parmi $(\mathbf{m}_1, \mathbf{m}_2)$; s'il s'agit de $\mathbf{m}_2$, on vérifie si la distance qui le sépare de $\mathbf{m}_1$ est bonne (entre d_1 et d_2) auquel cas on a trouvé le motif composite ; si la distance n'est pas bonne, il faut rechercher le prochain motif $\mathbf{m}_1$ et repartir de (**) ; s'il s'agissait du motif $\mathbf{m}_1$ alors on repart de (**). Connaissant les lois du temps d'attente avant la prochaine occurrence de $\mathbf{m}_1$ et du temps d'attente avant l'occurrence d'un des motifs $(\mathbf{m}_1, \mathbf{m}_2)$ [5], [10], on en déduit la loi du temps d'attente du motif composite. Cette approche n'est cependant valable que pour des motifs composites à 2 motifs.

[7] $P(A|B)$ désigne la probabilité de l'événement A sachant l'événement B.

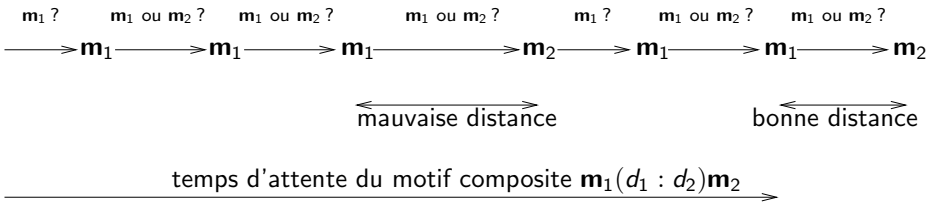

FIG. 1. Décomposition du temps d'attente avant l'occurrence du motif composite $\mathbf{m}_1(d_1 : d_2)\mathbf{m}_2$ comme la somme de temps d'attente indépendants entre les occurrences de $\mathbf{m}_1$ et $\mathbf{m}_2$.

– L'approche précédente peut se reformuler en termes de construction d'un processus semi-markovien dans lequel les états du processus correspondent aux différentes étapes nécessaires pour obtenir une occurrence du motif composite [12]. Pour 2 motifs il n'y a que 3 états : (*état 1*) le motif $\mathbf{m}_1$ est atteint (l'état suivant peut être l'état 1 ou l'état 2), (*état 2*) le motif $\mathbf{m}_2$ est atteint mais à la mauvaise distance de $\mathbf{m}_1$ (dans ce cas, l'état suivant sera l'état 1) et enfin (*état 3*) le motif $\mathbf{m}_2$ est atteint à la bonne distance de $\mathbf{m}_1$ (cet état est un état absorbant du processus semi-markovien). Si on reprend l'exemple de la figure 1, la succession des états serait la suivante : $\rightarrow 1 \rightarrow 1 \rightarrow 1 \rightarrow 2 \rightarrow 1 \rightarrow 1 \rightarrow 3$. Le temps d'attente du motif composite est donc égal au temps d'absorbance du processus semi-markovien[8] dont on sait calculer les probabilités de transition et les lois des temps de séjour dans chaque état grâce aux lois des temps d'attente entre motifs simples. Cette approche est facilement généralisable à un nombre quelconque de motifs simples.

3. Conclusion

Le lecteur intéressé par plus de détails mathématiques pourra se reporter au livre [7] ou aux chapitres d'ouvrages [3] et [9].

4. Références

[1] NUEL, G. (2004). LD-SPatt : Large Deviations Statistics for Patterns on Markov Chains. *Journal of Computational Biology*, **11**, 1023–1033.

[2] PRUM, B., RODOLPHE, F. and TURCKHEIM, À. (1995). Finding words with unexpected frequencies in DNA sequences, *Journal of the Royal Statistical Society series B*, **57**, 205–220.

[3] REINERT, G., SCHBATH, S. and WATERMAN, M. (2005). *Applied Combinatorics on Words*. volume 105 of *Encyclopedia of Mathematics and its Applications*, chapter Statistics on Words with Applications to Biological Sequences. Cambridge University Press.

[4] ROBIN, S. and DAUDIN, J.-J. (1999). Exact distribution of word occurrences in a random sequence of letters, *J. Appl. Prob.* **36,** 179–193.

[5] ROBIN, S. and DAUDIN, J.-J. (2001). Exact distribution of the distances between any occurences of a set of words, *Ann. Inst. Statist. Math.* **36,** 895–905.

[8] Le processus est semi-markovien dans le sens où les lois des temps de séjour dans chaque état ne sont pas géométriques, contrairement au cas markovien.

[6] Robin, S., Daudin, J.-J., Richard, H., Sagot, M.-F. and Schbath, S. (2002). Occurrence probability of structured motifs in random sequences, *Journal of Computational Biology*, **9**, 761–773.
[7] Robin, S., Rodolphe, F. and Schbath, S. (2003). *ADN, mots et modèles*. BELIN.
[8] Schbath, S. (1995). Compound Poisson approximation of word counts in DNA sequences, *ESAIM : Probability and Statistics*, **1**, 1–16.
[9] Schbath, S. and Robin, R. (2009). *Scan Statistics – Methods and Applications*. (J. Glaz, I. Pozdnyakov, and S. Wallenstein, ed.), chapter How can pattern statistics be useful for DNA motif discovery ? Statistics for Industry and Technology. Birkhauser.
[10] Stefanov, V.T. (2003). The intersite distances between pattern occurrences in strings generated by general discrete- and continuous-time models : an algorithmic approach, *J. Appl. Prob.* **40,** 881–892.
[11] Stefanov, V.T., Robin, S., and Schbath, S. (2007). Waiting times for clumps of patterns and for structured motifs in random sequences. *Discrete Appl. Math.* **155**, 868–880.
[12] Stefanov, V., Robin, S. and Schbath, S. (2011). Occurrence of structured motifs in random sequences : Arbitrary number of boxes. *Discrete Appl. Math.* **159**, 826–831.
[13] Sandve, G.K. and Abul, O. and Drablos, F. (2008). Compo : composite motif discovery using discrete models. *BMC Bioinformatics*, **9** :527.

Segmentation pour l'analyse de puces CGH

Emilie Lebarbier [1] et Franck Picard [2]

Chaque espèce possède un nombre caractéristique de copies des chromosomes : chez la grande majorité des êtres vivants, comme chez l'homme, chaque chromosome est présent en deux copies (organisme diploïde) mais d'autres organismes peuvent être polyploïdes (ayant plus de 2 copies de chaque chromosome), comme par exemple la pomme de terre ou l'huître avec 4 copies. Une déviation de ce nombre de copies par rapport au nombre normal pour l'espèce (surnuméraire ou manquante) entraîne de ce fait un déséquilibre du nombre de copies des gènes, pouvant être à l'origine de maladies majeures. Un exemple classique chez l'humain est la trisomie 21, qui se caractérise, comme son nom l'indique, par la présence de 3 copies du chromosome 21. La détection de ces défauts chromosomiques est donc un élément majeur dans l'établissement du diagnostic et/ou du traitement de la pathologie. Si la détection se place à l'échelle du chromosome, elle est rendue possible grâce au traditionnel caryotype. Cependant, la perte ou le gain peut ne toucher qu'une portion du chromosome. Et lorsque cette anomalie chromosomique est de très petite taille, elle peut passer inaperçue. C'est en 1992 que l'étude de ces « petites » anomalies connaît un essor considérable grâce à la mise au point d'une nouvelle technique, l'hybridation génomique comparative ou CGH. La résolution a été nettement améliorée par l'utilisation de la technologie des microarrays (microarrays CGH ou puces CGH) permettant la détection d'aberrations d'environ 50kb. Après avoir été exclusivement appliquées à l'étude de la génomique du cancer, les microarrays CGH sont aujourd'hui utilisées dans d'autres études de génétique humaine.

[1] UMR Agroparistech/INRA MIA 518.
[2] LBBE, UMR CNRS 5558 Université Lyon 1 Villeurbanne, France.

L'objectif des expériences de microarrays CGH est donc de détecter et de cartographier des aberrations chromosomiques à l'échelle du génome, en une seule expérience. Son principe consiste à compter les variations du nombre de copies de séquences d'ADN entre deux échantillons d'ADN génomiques (un échantillon test et un échantillon de référence). Ce nombre n'étant pas mesurable directement, la technologie des microarrays est utilisée (une description de cette technologie est donnée page 5 du chapitre d'introduction) : les deux échantillons sont marqués par deux molécules de couleur différente. Le mélange est ensuite déposé sur une lame sur laquelle sont fixées des séquences d'ADN du génome de l'organisme étudiée (appelées séquences cibles). Les séquences ADN colorées vont alors « s'apparier » avec leur séquence complémentaire déposée sur cette lame et autant de fois que la séquence est présente dans le mélange. Ensuite de l'excitation des molécules de couleurs, sont récupérées deux intensités (une par couleur). Ainsi la valeur de ces intensités va refléter l'abondance des séquences cibles dans chacun des échantillons. Afin d'en avoir une représentation visuelle, le log-ratio de l'intensité de l'échantillon test par rapport à celle de référence est calculé pour chaque séquence cible puis ordonné selon la position physique de ces séquences sur le génome (position connue). On obtient alors ce que l'on appelle un profil CGH.
Pour l'étude de syndrômes humains, l'échantillon de l'individu sain est pris comme référence et celui de l'individu malade comme test. L'homme étant diploïde, le nombre de copies de chaque séquence dans l'échantillon de référence est toujours 2. Dans la mesure où la quantité d'intérêt est un nombre relatif de copies de séquences d'ADN, les log-ratios devraient appartenir à un ensemble de valeurs prédéfinies : par exemple la perte d'une séquence dans l'échantillon test (une délétion) devrait montrer un rapport de $\log(1/2)$ alors que le gain (une amplification) devrait montrer un rapport de $\log(3/2)$. C'est ce que montre la figure 1 (Gauche). Notons que l'amplification peut être multiple (2, 3 voire 10 copies) et la délétion peut apparaître sur les deux chromosomes. Cependant, le signal obtenu en pratique est loin d'être discret (cf. figure 1 (Droite)). Cette variabilité, induite non seulement par la variabilité naturelle (nature des génomes étudiés) mais aussi par l'expérience, rend très compliquée une analyse « à la main ». La statistique s'avère alors essentielle pour pouvoir tenir compte de cette variabilité et extraire l'information biologique sous-jacente. Comme on peut le voir sur la figure 1 (Droite), la spécificité des données de microarrays CGH est qu'elles sont spatialement ordonnées. En effet, elles se présentent le long du génome comme une succession de segments représentant des régions sur le génome dont les séquences partagent en moyenne le même nombre de copies. La détection et la localisation de ces régions permettent alors une interprétation biologique du signal : des segments de moyenne positive s'interpréteront comme des régions amplifiées sur le génome de l'échantillon test et des segments de moyenne négative comme des régions délétées.

Les outils statistiques naturels pour cette problématique sont les méthodes de segmentation dont l'objectif général est de détecter les changements abrupts qui apparaissent dans certaines caractéristiques d'un signal. Dans le cadre particulier des données de microarrays CGH, il s'agira de déterminer : (i) quelles sont les caractéristiques soumises aux changements (c'est l'étape de modélisation) et (ii) combien il y a de changements et où ils sont (c'est l'étape d'inférence statistique). La segmentation est un sujet important en statistique et fait l'objet d'une recherche

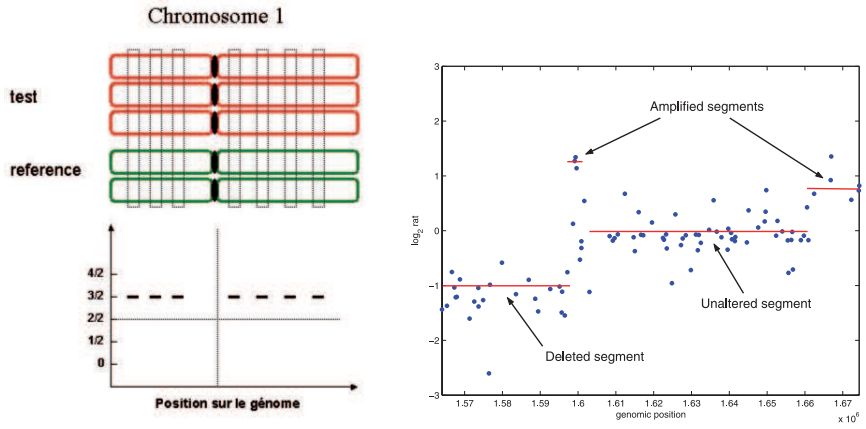

FIG. 1. Gauche : profil CGH théorique. Droite : profil CGH réel et interprétation. Ces figures sont extraites de [7].

intensive depuis plusieurs années. Cet intérêt est largement motivé par les nombreuses applications pratiques dans des domaines aussi variés que la biologie, la finance, la climatologie,... De nombreuses méthodes de segmentation ont été proposées pour l'analyse des données de microarrays CGH. Dans ce chapitre, nous en présentons une en particulier [7], qui s'est révélée être l'une des plus efficaces pour l'analyse de ce type de données [4].

1. Quel modèle pour les microarrays CGH ?

Il s'agit de déterminer un modèle décrivant aussi bien que possible le processus biologique. Ce processus peut être décrit par une fonction constante par morceaux : il existe une partition du génome notée $I_1, \ldots, I_K$ telle que les niveaux, valant μ_t à la position x_t et représentant le vrai log-ratio, sont constants à l'intérieur d'un intervalle et différents d'un segment à l'autre. On note μ_k la valeur du niveau au sein du segment I_k. Ces intervalles sont délimités par des instants, notés abusivement $t_1 < t_2 < \ldots < t_{K-1}$, appelés instants de rupture. Le signal observé n'est alors qu'une version bruitée de ce vrai signal : le log-ratio observé à la position x_t est défini par une variable aléatoire Y_t (pour $t = 1, \ldots, n$) telle que

$$Y_t = \mu_k + E_t \quad \text{si } t \in I_k.$$

La variabilité induite par l'expérience est prise en compte au travers de la variable aléatoire E_t. Ces variables sont supposées indépendantes et être issues d'une distribution gaussienne de moyenne nulle et de variance σ^2. Cela revient à supposer que la distribution des Y_t est une gaussienne de moyenne μ_k et de variance σ^2 ($\mathcal{N}(\mu_k, \sigma^2)$) si t appartient au segment I_k. Ce modèle est noté $\mathcal{M}_m$. Notons cependant qu'un autre modèle pourrait être envisagé : à la fois la moyenne et la variance peuvent être sujettes aux changements. Dans ce cas, le modèle noté $\mathcal{M}_{mv}$ s'écrit : $Y_t \sim \mathcal{N}(\mu_k, \sigma_k^2)$.

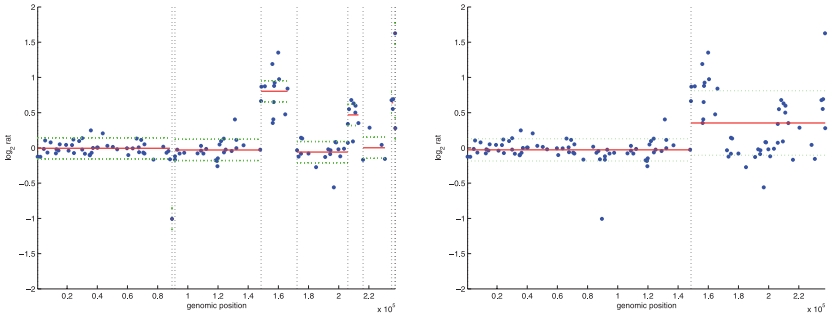

FIG. 2. Segmentations pour le modèle $\mathcal{M}_m$ (Gauche), pour le modèle $\mathcal{M}_{mv}$ (Droite)

Pour savoir quel modèle est le plus adapté à l'analyse des données CGH, il est non seulement important d'avoir une bonne connaissance des données mais aussi de bien connaître le comportement des deux modèles. Par définition, le modèle $\mathcal{M}_{mv}$ autorise des variances différentes selon les segments. Ainsi avec ce modèle de petites régions successivement de niveaux différents pourront être regroupées en une longue région de grande variance, alors qu'elles seront chacune considérées comme un segment avec le modèle $\mathcal{M}_m$, comme l'illustre la figure 2. Cet argument milite en faveur de ce dernier modèle puisqu'il permet de mettre en exergue plus de régions, certes des régions pouvant être issues d'artefacts techniques mais aussi des régions d'intérêt biologique. On peut aussi noter une récente procédure de segmentation proposée par [1] dans le cadre de changements dans la moyenne mais sans contrainte sur la variance.

2. Comment obtenir la meilleure segmentation ?

Pour un nombre K de segments, une segmentation est décrite par les instants de rupture t_k et les moyennes μ_k. Ainsi rechercher la « meilleure » segmentation en K segments se réduit à déterminer les « meilleures » valeurs de ces paramètres. À cet effet, il est nécessaire de définir un critère de qualité des segmentations. Un critère classique en statistique est le critère des moindres carrés qui permet de mesurer l'écart existant entre la segmentation proposée et le signal observé. Il s'écrit ici :

$$J_K = \sum_{k=1}^{K} \sum_{t \in I_k} (Y_t - \mu_k)^2. \tag{1}$$

Trouver les valeurs des μ_k et t_k qui minimisent ce critère revient donc à chercher la segmentation en K segments qui s'ajuste le mieux aux données. La particularité des paramètres des instants de rupture t_k est qu'ils sont discrets (les moyennes étant continues). Ainsi la minimisation du critère portant sur ces paramètres nécessite, dans le cas d'une recherche exhaustive, l'exploration de C_{n-1}^{K-1} configurations, rendant impossible l'utilisation d'un algorithme naïf. Grâce à la propriété d'additivité de J_K en K, la segmentation optimale peut être obtenue en un temps raisonnable par un algorithme de Programmation Dynamique [2], emprunté à la théorie des graphes.

Cet algorithme repose sur le principe d'optimalité énoncé par le mathématicien Richard Bellman : « tout chemin optimal est formé de sous-chemins optimaux ». Cet algorithme a été récemment amélioré [10] permettant l'analyse de signaux de très grande taille, comme c'est le cas des nouvelles données de microarrays CGH qui comptent au moins un million de points.

Deux questions se posent alors. (i) Comment choisir le nombre K de segments ? (ii) Quelle confiance peut-on accorder à la segmentation obtenue ? Les réponses à ces questions ne sont pas si simples. Là encore le problème principal est la nature discrète des instants de rupture, qui ne permet pas d'utiliser les outils classiques de statistique.

Choix du nombre de segments K.

On dispose donc de la meilleure segmentation des données en K segments. Cependant, en pratique ce nombre n'est pas connu. Et même si l'on dispose d'informations a priori sur les données, il est difficile de se le fixer à l'avance, il s'agit donc de le déterminer. La qualité de cette meilleure segmentation est naturellement mesurée par la valeur du minimum de J_K en μ_k et t_k à K fixé, et noté $\hat{J}_K$. $\hat{J}_K$ traduit l'ajustement du modèle aux données. Ce terme diminue avec le nombre de segments K : plus il y a de segments, plus la segmentation résultante sera proche des données. On est alors tenté de choisir le nombre maximal de segments pour avoir le meilleur ajustement. Dans ce cas, la segmentation résultante coïncide complètement avec les données, ce qui est sans intérêt. On a plutôt envie de choisir une segmentation suffisamment bien ajustée aux données sans être trop complexe, c'est-à-dire avec un nombre raisonnable de segments, afin d'obtenir une représentation des données suffisamment informative sans qu'elle soit difficilement interprétable. La traduction mathématique de ce compromis s'obtient en ajoutant au critère des moindres carrés, une fonction notée *pen* et appelée pénalité :

$$\hat{J}_K + pen(K).$$

Cette pénalité doit donc être une fonction qui reflète la complexité de la segmentation et augmenter avec elle. Le nombre de segments est ensuite choisi en minimisant ce critère, appelé critère pénalisé. Bien sûr toute la difficulté réside dans le choix d'une fonction de pénalité judicieuse. Bien que depuis ces dernières années de nombreux travaux aient été menés sur ce sujet ([5], [3], [6], [12], ...), il n'existe pas de critère universellement meilleur. La sélection du nombre de segments reste un problème ouvert.

Qualité de la segmentation obtenue.

La méthode de segmentation précédente propose au final une segmentation particulière. Mais à quel point peut-on avoir confiance en cette segmentation ? À quel point est-on sûr de la rupture localisée en t ? Cette dernière question se traduit statistiquement par « quelle est la probabilité qu'une rupture soit localisée en t ? ». Plus la probabilité est proche de 1, plus forte sera la certitude sur son existence. Une approche possible est l'approche bayésienne. Dans ce cadre, les paramètres (les μ_k et t_k) sont vus comme des variables aléatoires, donnant alors un sens à la probabilité qu'une rupture soit localisée en t ou comprise dans un intervalle. Le calcul exact de ces probabilités nécessite l'exploration de l'espace des segmentations (plus ou moins restreint selon la probabilité d'intérêt), qui, comme évoqué précédemment, pose d'importantes difficultés algorithmiques. Ces difficultés

peuvent être contournées pour certains modèles de segmentations [11], comme par exemple pour le modèle $\mathcal{M}_{mv}$.

3. Vers une nouvelle dimension

Aujourd'hui l'avancée importante de la technologie des expériences de microarrays offre la possibilité d'analyser simultanément plusieurs échantillons biologiques. L'arrivée de ces nouvelles données a généré de nouvelles questions, comme l'analyse jointe d'aberrations chromosomiques sur un ensemble de profils. Une telle analyse offre d'une part l'opportunité de corriger des artefacts techniques, qui sont partagés par tous les signaux issus d'une même lame, et d'autre part d'intégrer des informations sur chaque échantillon, comme des informations cliniques dans l'étude des cancers. Le passage au niveau multi-profils pose évidemment de nouvelles questions de modélisation statistique et amène de nouveaux défis algorithmiques [8].

4. Références

[1] S. Arlot and A. Celisse, *Segmentation of the mean of heteroscedastic data via cross-validation*, Statistics and Copmputing, in prints (technical report arXiv :0902.3977v2, 2009) (2010).

[2] R.E. Bellman and S.E. Dreyfus, *Applied dynamic programming*, Princeton University Press, 1962.

[3] L. Birgé and P. Massart, *Minimal penalties for gaussian model selection*, Probability Th. and Related Fields **138** (2007), 33–73.

[4] W.R. Lai, M.D. Johnson, R. Kucherlapati, and P. J. Park, *Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data*, Bioinformatics **0** (2005), n° 0, 1–8.

[5] M. Lavielle, *Using penalized contrasts for the change-point problem*, Signal Processing **85** (2005), n° 8, 1501–1510.

[6] E. Lebarbier, *Detecting multiple change-points in the mean of gaussian process by model selection*, Signal Processing **85** (2005), 717–736.

[7] F. Picard, *Process segmentation/clustering. application to the analysis of array CGH data*, Ph.D. thesis, 2005.

[8] F. Picard, E. Lebarbier, M. Hoebeke, G. Rigaill, B. Thiam, and Robin, *Joint segmentation, calling and normalization of multiple CGH profiles*, Biostatistics **12** (2011), n° 3, 413–428.

[9] F. Picard, S. Robin, M. Lavielle, C. Vaisse, and J.-J. Daudin, *A statistical approach for array CGH data analysis*, BMC Bioinformatics **6** (2005), n° 27, 1, `www.biomedcentral.com/1471-2105/6/27`.

[10] G. Rigaill, *Pruned dynamic programming for optimal multiple change-point detection*, Tech. report, arXiv :1004.0887v1, 2010.

[11] G. Rigaill, E. Lebarbier, and S. Robin, *Exact posterior distributions over the segmentation space and model selection for multiple change-point detection problems*, Tech. report, arXiv :1004.4347, 2011.

[12] N. R. Zhang and D. O. Siegmund, *A modified Bayes information criterion with applications to the analysis of comparative genomic hybridization data*, Biometrics **63** (2007), n° 1, 22–32.

Tests multiples en génomique

Pierre Neuvial[1]

1. Recherche de gènes différentiellement exprimés

L'émergence récente de problèmes de grande dimension, c'est-à-dire pour lesquels le nombre p de variables excède le nombre n d'observations – et ce parfois de plusieurs ordres de grandeur, nécessite des développements majeurs en statistique. En effet, les outils mathématiques classiques pour traiter des questions usuelles comme les problèmes de tests d'hypothèses ou de classification (supervisée ou non supervisée) ont été développés dans des contextes où $p < n$, et ne sont donc pas applicables tels quels en grande dimension.

Nous nous intéressons dans ce chapitre à la question des *tests multiples en génomique*, que nous introduisons grâce à l'exemple de la recherche de gènes dont le niveau d'expression (c'est-à-dire le niveau d'activité) diffère entre deux groupes de patients à partir de données de puces à ADN. On parle de *gènes différentiellement exprimés*. Notons qu'il existe des applications de la théorie des tests multiples dans tous les domaines faisant intervenir des données de grande dimension, non seulement en génomique mais aussi par exemple en imagerie médicale ou en astronomie.

En guise d'illustration, nous utilisons un des premiers jeux de données publiés : les données de Golub [2]. Il s'agit d'une collection de $n = 38$ expériences de puces à ADN effectuées sur des prélèvements sanguins de deux groupes de 11 et 27 patients atteints de deux types de leucémies (cancers du sang). Chacune de ces expériences fournit une mesure du niveau d'expression des mêmes $p = 3051$ gènes chez un des 38 patients. Un des objectifs de l'étude est l'identification de gènes différentiellement exprimés entre les deux types de leucémies. Dans les jeux de données plus récents, p est généralement de l'ordre de plusieurs dizaines de milliers, tandis que n dépasse rarement une centaine d'observations.

Ces données de puces à ADN sont représentées dans la figure 1. L'asymétrie de la matrice (a) est typique des données de grande dimension : ses lignes correspondent aux 3051 gènes et ses colonnes aux 38 patients. La matrice (b) est une extraction des 367 lignes de la matrice (a) dont le niveau d'expression diffère le plus fortement (en un sens précisé à la Section 2) entre les deux groupes, c'est-à-dire des gènes les plus différentiellement exprimés. L'objectif de ce chapitre est d'expliquer, en s'appuyant sur cet exemple, les enjeux statistiques sous-tendus par l'extraction d'un tel sous-ensemble d'hypothèses.

[1] Laboratoire Statistique et Génome, Université d'Évry Val d'Essonne, UMR CNRS 8071 – USC INRA.

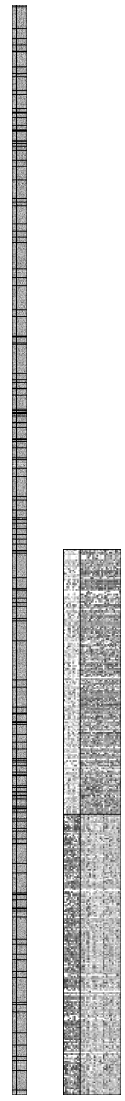

FIG. 1. Recherche de gènes différentiellement exprimés en grande dimension : matrices d'expression, avec les individus (patients) en colonne, et les variables (gènes) en ligne. La matrice des données originales (a) est de dimension 3051 x 38 ; la matrice (b), de dimension 367 x 38, est un exemple de résultat : elle correspond à l'extraction des 367 lignes de (a) marquées en noir. Le niveau d'expression d'un gène est représenté par une couleur allant du noir (faible expression) au blanc (forte expression).

Si la théorie des tests statistiques permet, pour un gène donné, de quantifier la différence entre les niveaux d'expression des deux groupes, et même d'associer un degré de confiance à cette quantification, elle ne permet pas en revanche d'associer un tel degré de confiance à un ensemble de gènes testés simultanément. C'est l'objet essentiel de la théorie des tests multiples, qui inclut d'une part la définition de mesures de risques adaptées à un ensemble de tests, et d'autre part la recherche d'algorithmes permettant de contrôler ces risques ainsi que de conditions réalistes garantissant l'applicabilité de ces procédures en pratique. Pour une introduction plus complète à la théorie des tests multiples, nous renvoyons à un article récent du *Journal de la Société Française de Statistique* [4].

2. Tests

Un test statistique est une procédure permettant de valider ou d'invalider une hypothèse de travail. On s'intéresse ici au cas particulier où on cherche à identifier une différence entre les niveaux observés d'une variable pour deux groupes d'individus. Dans notre exemple, considérons un gène (parmi p). La variable observée est le niveau d'expression de ce gène, et les individus sont des patients atteints de deux types de leucémies. Elle correspond à une ligne donnée dans les matrices de la figure 1. Les ingrédients classiques pour la mise en place d'un test sont les suivants.

(1) Un modèle probabiliste, qui formalise les hypothèses faites sur la distribution de la variable observée. Par exemple, on pourra supposer que les niveaux d'expression du gène suivent une loi gaussienne d'espérance (c'est-à-dire de moyenne) μ_1 pour le groupe 1 et μ_2 pour le groupe 2.

(2) Une hypothèse de travail, qu'on cherche à valider ou invalider, qu'on appelle « hypothèse nulle » et qu'on note $\mathcal{H}_0$. Dans notre exemple, il s'agit d'une définition mathématique de l'absence de différences réelles, comme l'égalité des moyennes μ_1 et μ_2.

(3) Une règle de décision, qui associe à un jeu de données (ici, 11 observations issues du premier groupe et 27 issues du second) le rejet ou l'acceptation de l'hypothèse nulle. Par construction, elle garantit que la probabilité de rejeter l'hypothèse nulle alors qu'elle est vraie ne dépasse pas une valeur pré-spécifiée. Cette valeur est appelée *niveau* du test, et sera notée α. On appelle *probabilité critique* associée à un jeu de données le plus petit niveau de test permettant de rejeter l'hypothèse nulle.

La contribution principale de la statistique se situe à l'étape de construction d'une règle de décision. Dans notre exemple, l'approche la plus classique est le test de Student [3], dont la mesure est le ratio entre une estimation de la différence des moyennes des deux groupes, et une estimation de la variabilité. Sous l'hypothèse nulle, c'est-à-dire si $\mu_1 = \mu_2$, ce ratio suit une loi de probabilité connue, ce qui induit une règle de décision. Les 367 gènes de la figure 1(b) correspondent aux variables de plus grands ratios parmi les 3051 gènes initiaux.

3. Tests multiples

La « recette » de la section précédente permet de tester *une* hypothèse à un niveau α donné. Pour un gène donné, on sait donc décider s'il existe une

différence d'expression réelle entre deux groupes d'échantillons, avec un risque d'erreur contrôlé. Considérons maintenant la situation où l'on effectue simultanément le test de p hypothèses au niveau α. Chaque test a 4 issues possibles, selon que l'hypothèse nulle $\mathcal{H}_0$ est rejetée ou non, et selon qu'elle est vraie ou fausse (Tableau 1).

	$\mathcal{H}_0$ acceptée (négatifs)	$\mathcal{H}_0$ rejetée (positifs)	Ensemble
$\mathcal{H}_0$ vraie	VN	FP	p_0
$\mathcal{H}_0$ fausse	FN	VP	$p - p_0$
Ensemble	$p - R$	R	p

TAB. 1. Ventilation de p hypothèses testées en termes de nombres de vrais négatifs (VN), faux négatifs (FN), faux positifs (FP), vrais positifs (VP). R est le nombre total d'hypothèses rejetées, et p_0 le nombre (inconnu) d'hypothèses nulles vraies.

Notons p_0 le nombre total d'hypothèses nulles vraies (gènes non différentiellement exprimés). Le nombre FP de faux positifs, c'est-à-dire d'hypothèses rejetées à tort, est en moyenne $p_0\alpha$. Par conséquent, en supposant qu'aucun gène n'est réellement différentiellement exprimé (c'est-à-dire que $p_0 = p$), on attend environ 150 faux positifs si l'on effectue le test de $p = 3051$ gènes au seuil $\alpha = 5\%$.

3.1. Taux d'erreurs par famille : FWER

Pour limiter cette inflation du nombre de faux positifs, il est naturel de diminuer le niveau de chacun des tests individuels, dans le but de contrôler un risque global, associé aux p tests. Ainsi, si chaque test est effectué au niveau α', il est possible de contrôler la probabilité qu'il y ait au moins un faux positif, que l'on appelle taux d'erreur par famille, ou FWER pour Family-Wise Error Rate. En effet, on a par sous-additivité de la mesure de probabilité

$$\text{FWER} \leqslant \sum_{\{i/\mathcal{H}_0 \text{ est vraie pour } i\}} \mathbb{1}_{\{\mathcal{H}_0 \text{ est rejetée pour } i\}}$$

Chaque rejet étant un événement de probabilité α' par définition, on obtient $\text{FWER} \leqslant p_0\alpha'$. Ainsi, effectuer chaque test au niveau $\alpha' = \alpha/p$ assure que la probabilité qu'il y ait au moins un faux positif est majorée par α (car $p_0 \leqslant p$). Cette procédure de test multiples, dite de Bonferroni, est peu utilisée dans les applications génomiques. En effet, le grand nombre p d'hypothèses testées la rend souvent trop stringente le seuil α/p étant fréquemment si petit qu'aucune hypothèse n'est déclarée significative. Elle fonctionne cependant bien dans l'exemple des données de Golub, où l'on obtient 98 rejets au seuil $\alpha = 5\%$.

3.2. Taux de fausses découvertes : FDR

La mesure de risque la plus couramment utilisée aujourd'hui dans les problématiques de tests multiples est le False Discovery Rate (FDR) [1]. Le FDR est la proportion attendue de faux positifs parmi les hypothèses rejetées, qui correspond donc à la proportion de gènes considérés à tort comme différentiellement exprimés parmi les gènes sélectionnés, soit, avec les notations du Tableau 4.1 :

$$\mathrm{FDR} = \mathbb{E}\left[\frac{\mathrm{FP}}{R \vee 1}\right].$$

Benjamini et Hochberg ont prouvé qu'une procédure (due à Simes [5]) permet de contrôler le FDR dans le cas où les hypothèses nulles vraies sont indépendantes. Soient $(X_i)_{1\leqslant i\leqslant p}$ les probabilités critiques associées à p tests d'hypothèses. On note

$$\hat{k} = \max\{1 \leqslant i \leqslant p : X_{(i)} \leqslant \alpha i/m\},$$

où $(X_{(i)})_{1\leqslant i\leqslant p}$ est un réarrangement croissant de $(X_i)_{1\leqslant i\leqslant p}$. La procédure de Simes rejette toutes les hypothèses dont les probabilités critiques sont en deçà de $X_{(\hat{k})}$. Avec les données de Golub, on obtient 367 rejets pour un niveau FDR cible de 5% : ce sont les gènes de la figure 1(b).

Le risque FDR et la procédure de Simes sont rapidement devenus des standards pour les problèmes de tests multiples. Parallèlement, cette mesure de risque a engendré de nombreux développements statistiques récents en théorie des tests multiples. Nous en évoquons trois très brièvement, et renvoyons à [4] pour plus de détails.

– L'amélioration des procédures de contrôle du FDR via l'estimation de la proportion $\pi_0 = p_0/p$ d'hypothèses nulles vraies. En effet, la procédure de Simes appliquée à un niveau α garantit en fait un contrôle du FDR au niveau $\pi_0\alpha$, donc strictement inférieur au niveau prescrit dès que $\pi_0 < 1$.

– L'étude de l'impact de la dépendance entre les hypothèses testées sur le contrôle du FDR, afin de rendre les hypothèses faites sur cette dépendance plus proches des structures de dépendance observées en pratique, notamment dans les données génomiques.

– La définition d'autres mesures de risque inspirées par des limitations du contrôle du FDR. En particulier, le FDR est la valeur moyenne de la proportion (aléatoire) de faux positifs parmi les hypothèses rejetées, qui est notée FDP pour False Discovery Proportion. Par conséquent, contrôler le FDR ne renseigne pas sur les fluctuations de cette proportion ; il peut donc être intéressant de contrôler d'autres caractéristiques de la *distribution* du FDP, notamment ses queues de distribution.

4. De nouveaux enjeux méthodologiques

Outre les développements mathématiques que nous venons d'évoquer, la théorie des tests multiples doit faire face à de nouveaux enjeux méthodologiques, motivés par les possibilités d'exploiter de nouveaux types de données génomiques. Un premier axe est l'utilisation d'informations *a priori* sur les hypothèses à tester, notamment sur leur structure de dépendance. Dans le cas de la recherche de gènes différentiellement exprimés, les biologistes disposent souvent, outre des données

de puces à ADN, d'informations sur les réseaux d'interactions régissant les relations fonctionnelles entre les gènes (voir le chapitre sur les réseaux). Plutôt que de procéder en deux étapes : analyse différentielle pour sélectionner une liste de gènes, puis interprétation des résultats à la lumière des connaissances fonctionnelles, il paraît naturel de tester directement l'expression différentielle de réseaux de gènes (connus). Ceci implique d'une part la définition de procédures de tests appropriées pour un réseau donné, et d'autre part le développement de procédures de tests multiples adaptées.

Un autre axe important est l'arrivée récente de données de séquençage à haut débit, notamment pour la quantification des niveaux d'expression des gènes (RNA-seq). Ces données sont d'encore plus grande dimension, puisqu'on lit désormais des *millions* de séquences en une seule expérience, et leur nature diffère fondamentalement des données de puces à ADN car ce sont des données de comptage.

5. Références

[1] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate : a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B*, 57(1) : 289–300, 1995.

[2] T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, and E.S. Lander. Molecular classification of cancer : class discovery and class prediction by gene expression monitoring. *Science*, 286(5439) :531–537, October 1999.

[3] W.S. Gosset. The probable error of a mean. *Biometrika*, 6(1) :1–25, 1908.

[4] E. Roquain. Type I error rate control for testing many hypotheses : a survey with proofs. *J. Soc. Fr. Stat.*, To appear.

[5] R.J. Simes. An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, 73(3) :751–754, 1986.

Réseaux biologiques

Julien Chiquet[1]

L'essor d'internet et des réseaux sociaux a popularisé les graphes comme outil de représentation de relations entre variables dans tous les champs d'application. En biologie des systèmes, on les retrouve à divers niveaux de description des mécanismes, ainsi i) les réseaux métaboliques (graphes entre réactions et produits de réaction) ; ii) les réseaux protéines/protéines (graphes de proximité des séquences protéiques) ; iii) les réseaux de régulation (graphes des interactions du produit d'un gène sur la transcription d'autres gènes) ; ou iv) les réseaux de co-expression (graphes des corrélations entre niveau d'expressions de gènes).

Si l'analyse des réseaux biologiques a beaucoup emprunté aux télécommunications et aux sciences sociales, domaines où la théorie des graphes est solidement ancrée, les systèmes biologiques ont généré de nouvelles problématiques dues à l'extrême complexité des mécanismes en jeu, à leur instabilité et à la nature des sources de données associées (typiquement un très grand nombre de variables

[1] Laboratoire Statistique et Génome, Université d'Évry Val d'Essonne, UMR CNRS 8071 – USC INRA.

pour peu d'observations). Dès lors les statistiques se sont imposées pour gérer les incertitudes liées aux phénomènes et à l'afflux massif de données de grande dimension, rendant souvent prohibitive l'utilisation à grande échelle de modèles déterministes.

Cet article se veut une brève introduction à quelques outils de statistique pour l'étude des réseaux biologiques. Nous présentons un cadre de modélisation largement répandu chez les statisticiens et nous survolons deux visions du réseau suscitant de nombreux travaux en statistiques mathématiques et appliquées : la première considère le réseau biologique comme étant l'observation ; la seconde s'appuie sur les données omiques pour reconstruire les arêtes du réseau. Nous tâcherons d'insister sur les spécificités de ces outils liées aux contraintes d'ordre biologique et à l'échantillonnage.

1. Modélisation

Notations. Pour fixer les idées, on considère dans la suite uniquement les graphes non dirigés composés d'un ensemble fixe $\mathcal{P} = \{1, \dots, p\}$ de nœuds, correspondant aux molécules d'intérêt (gènes ou protéines, par exemple). La présence d'une arête entre deux nœuds i et j de $\mathcal{P}$ est définie par la variable aléatoire indicatrice $\Theta_{ij} = \mathbb{1}_{\{i \leftrightarrow j\}}$, ainsi $\mathbf{\Theta} = (\Theta_{ij})_{i,j \in \mathcal{P}^2}$ décrit la matrice (aléatoire) d'adjacence du graphe. Par convention, les nœuds ne sont pas connectés à eux-mêmes, c'est-à-dire que $\Theta_{ii} = 0$ pour tout $i \in \mathcal{P}$.

Partant d'une collection de réseaux connus, on souhaite naturellement proposer un modèle suffisamment souple pour reproduire les attributs statistiques dominants de cette collection, comme la distribution des degrés des nœuds : typiquement, les réseaux biologiques se caractérisent par la présence d'un grand nombre de nœuds à faible degré et d'un petit nombre au degré très élevé, appelés « hubs ». Une telle distribution peut être décrite par une loi de puissance, popularisée par la physique statistique (voir [2]). Elle s'avère cependant trop souvent restrictive pour une description fine d'autres caractéristiques des réseaux biologiques, telles leur forte hétérogénéité ou le faible nombre effectif d'arêtes parmi le nombre possible : le nombre d'arêtes est typiquement de l'ordre du nombre de nœuds.

Le modèle à « blocs stochastiques » [2] s'intègre bien à ces problématiques : il s'agit d'un modèle de graphe aléatoire issu des sciences sociales réintroduit indépendamment par de nombreux auteurs. On trouvera dans [4] une présentation complète et de récentes extensions. Ce modèle propose de distribuer les nœuds $\mathcal{P}$ en un ensemble $\mathcal{Q} = \{1, \dots, Q\}$ de classes latentes[3] de probabilités *a priori* $\boldsymbol{\alpha} = (\alpha_q)_{q \in \mathcal{Q}}$. Par la suite, on suppose pour simplifier Q connu et fixé. L'appartenance du nœud i à une classe est commodément décrite par le vecteur $\mathbf{Z}_i = (Z_{iq})_{q \in \mathcal{Q}}$ qui suit une loi multinomiale $\mathcal{M}(1, \boldsymbol{\alpha})$. Chacune des classes permet de représenter un groupe de variables dont le comportement dans le graphe est homogène, aussi bien entre elles que vis-à-vis des variables extérieures au groupe. On pense notamment à un ensemble de molécules impliquées dans les mêmes voies métaboliques. Cette propension à la connexion interne ou externe des nœuds d'un

[2] *Stochastic Bloc Model* en anglais.
[3] C'est-à-dire non observées.

groupe est décrite par une série de paramètres $\boldsymbol{\pi} = (\pi_{ql})_{q,l\in\mathcal{Q}}$, indiquant la probabilité pour qu'un nœud de la classe q se connecte avec un nœud de la classe l. La probabilité de présence d'une arête entre les nœuds i et j est définie *conditionnellement*[4] à leurs classes respectives selon une loi de Bernoulli :

$$\Theta_{ij} \mid \{Z_{iq}Z_{jl} = 1\} \sim \mathcal{B}(\pi_{ql}), \quad i \neq j. \tag{1}$$

Ce modèle s'avère particulièrement souple, permettant la génération d'une grande variété de topologies de graphes composés de sous-graphes bipartites, en communauté ou en étoile par exemple.

2. Analyse des réseaux aléatoires

Une première classe de problèmes consiste à considérer l'objet réseau $\boldsymbol{\Theta}$ comme étant l'observation. Ajuster un modèle à un graphe biologique est très utile à la découverte d'éléments structurants qui n'étaient pas directement visibles : les classes de nœuds correspondant à des classes de gènes peuvent permettre aux biologistes de formuler des hypothèses recoupant les connaissances disponibles souvent parcellaires. Par exemple, si dans un réseau de régulation un gène se trouve impliqué dans une structure en étoile, il est possible qu'il contrôle la traduction d'une protéine régulant la transcription d'autres gènes ; de même, l'existence d'une communauté de gènes fortement connectés suggère une implication des protéines sous-jacentes dans des voies métaboliques communes, parfois inconnues. Ainsi l'analyse de réseau permet la formulation d'hypothèses de travail pour les biologistes, voire leur validation dans certains cas, en vue d'isoler des structures ou motifs caractérisant une espèce ou une maladie.

Du point de vue statistique, l'ajustement du modèle à blocs stochastiques correspond à déterminer le paramètre $\boldsymbol{\beta}$ regroupant les proportions de classe et les probabilités de connexion, soit $\boldsymbol{\beta} = \{\boldsymbol{\alpha}, \boldsymbol{\pi}\}$. La méthode d'estimation du maximum de vraisemblance est la plus naturelle en statistique : elle consiste à déterminer les valeurs des paramètres les plus vraisemblables vis-à-vis des données, en maximisant la probabilité d'occurrence des données sous le modèle choisi. On note $\mathbb{P}_{\boldsymbol{\beta}}(\boldsymbol{\Theta})$ cette probabilité, correspondant à la loi du graphe observé sous les paramètres $\boldsymbol{\beta}$, inconnus. La log vraisemblance associée, notée ℓ, « renverse » le problème en considérant les observations comme étant un paramètre fixé et $\boldsymbol{\beta}$ la variable. Ainsi, on cherche à résoudre le problème

$$\arg\max_{\boldsymbol{\beta}} \ell_{\boldsymbol{\Theta}}(\boldsymbol{\beta}) = \arg\max_{\boldsymbol{\beta}} \mathbb{P}_{\boldsymbol{\beta}}(\boldsymbol{\Theta}).$$

La spécificité du problème courant tient au fait qu'une partie de l'information disponible dans les données est manquante : on observe le graphe $\boldsymbol{\Theta}$ alors que le modèle (1) requiert l'appartenance aux classes des nœuds via $\mathbf{Z} = \{\mathbf{Z}_i\}_{i\in\mathcal{P}}$. La log vraisemblance du modèle ne peut être écrite qu'en marginalisant sur $\mathbf{Z}$, la distribution des arêtes Θ_{ij} n'étant connue qu'à classes fixées :

$$\ell_{\boldsymbol{\Theta}}(\boldsymbol{\beta}) = \mathbb{P}_{\boldsymbol{\beta}}(\boldsymbol{\Theta}) = \log \sum_{\mathbf{Z}\in\mathcal{Z}} \mathbb{P}_{\boldsymbol{\beta}}(\boldsymbol{\Theta}, \mathbf{Z}) = \log \sum_{\mathbf{Z}\in\mathcal{Z}} \mathbb{P}_{\boldsymbol{\beta}}(\boldsymbol{\Theta}|\mathbf{Z})\mathbb{P}_{\boldsymbol{\beta}}(\mathbf{Z}).$$

[4] La loi conditionnelle est symbolisée par le symbole |.

Maximiser $\ell(\boldsymbol{\beta})$ s'avère impossible directement numériquement de par la présence de $Q^{|\mathcal{P}|}$ termes dans la sommation, correspondant à toutes les configurations possibles de $\mathbf{Z}$. Cette famille de problèmes, à savoir la maximisation de la vraisemblance d'un modèle à classes latentes, est bien connue en statistique depuis les années 1970 et se contourne à l'aide de la stratégie EM. Celle-ci repose sur la décomposition suivante de la log vraisemblance, utilisant la vraisemblance des données *complétées* par les variables latentes :

$$\ell_{\boldsymbol{\Theta}}(\boldsymbol{\beta}) = \log \mathbb{P}_{\boldsymbol{\beta}}(\boldsymbol{\Theta}, \mathbf{Z}) - \log \mathbb{P}_{\boldsymbol{\beta}}(\mathbf{Z}|\boldsymbol{\Theta}). \tag{2}$$

L'algorithme EM consiste à alterner une étape approchant la distribution conditionnelle de la classification $\mathbf{Z}$ pour une valeur fixe des paramètres $\boldsymbol{\beta}$ (i.e., le deuxième terme dans (2)) et une étape déterminant les valeurs des paramètres $\boldsymbol{\beta}$ maximisant la vraisemblance complétée (le premier terme dans (2)) calculée sous l'estimation courante de la distribution conditionnelle de $\mathbf{Z}$. La suite ainsi construite pour les paramètres $\boldsymbol{\beta}$ conduit à un maximum local de $\ell_{\boldsymbol{\Theta}}$.

Malheureusement, la stratégie EM échoue elle aussi dans le cas du modèle à blocs stochastiques car elle requiert le calcul explicite de $\mathbb{P}(\mathbf{Z}|\boldsymbol{\Theta})$, ce qui s'avère impossible de par la complexité des structures de dépendances entre variables dans le cas des modèles de graphes aléatoires. Pour résoudre ce problème, les statisticiens se sont tournés notamment vers les approches variationelles : elles consistent à maximiser une borne inférieure de la vraisemblance, obtenue en approchant la distribution $\mathbb{P}(\mathbf{Z}|\boldsymbol{\Theta})$ par une forme factorisable, rendant possible les calculs (voir par exemple [3]).

Diverses extensions du modèle à blocs stochastiques et des méthodes d'estimation associées ont mené à de nombreuses publications, souvent dues au soucis de cohérence avec la biologie. Nous pensons par exemple à la possibilité pour un nœud d'appartenir à plusieurs classes, c'est-à-dire, pour une protéine, d'être au carrefour de plusieurs voies métaboliques : on parle de classe *chevauchante*.

À titre d'exemple, nous proposons à la figure 1 une vue du modèle à blocs stochastiques à 5 classes ajusté au réseau de régulation d'*Escherichia coli*, organisme modèle en biologie. Les nœuds sont les gènes de l'organisme et les arêtes décrivent des relations de régulations (activation ou inhibition) induites par les protéines codées par chacun des gènes. On distingue une structure très forte, en particulier la présence de « hubs » et de communautés correspondant respectivement à des gènes régulateurs, appelés facteurs de transcription, et à des groupes formés de gènes tous régulés par un même autre.

3. Reconstruction

Le problème de la reconstruction de réseau, parfois abusivement appelé « inférence de réseau », est radicalement différent. Il consiste, à partir d'une source de données relatives aux variables de $\mathcal{P}$, à déterminer les arêtes du graphe, c'est-à-dire à déterminer les éléments non nuls et éventuellement leur intensité dans la matrice $\boldsymbol{\Theta}$: le graphe considéré est alors *valué*.

À cet effet, il convient de définir une mesure qui permette de rendre compte des interactions effectives entre les variables d'étude. Cette mesure dépend directement de la nature des données biologiques en jeu. Dans le cadre de l'analyse du transcriptome, un échantillon correspond typiquement au niveau d'expression

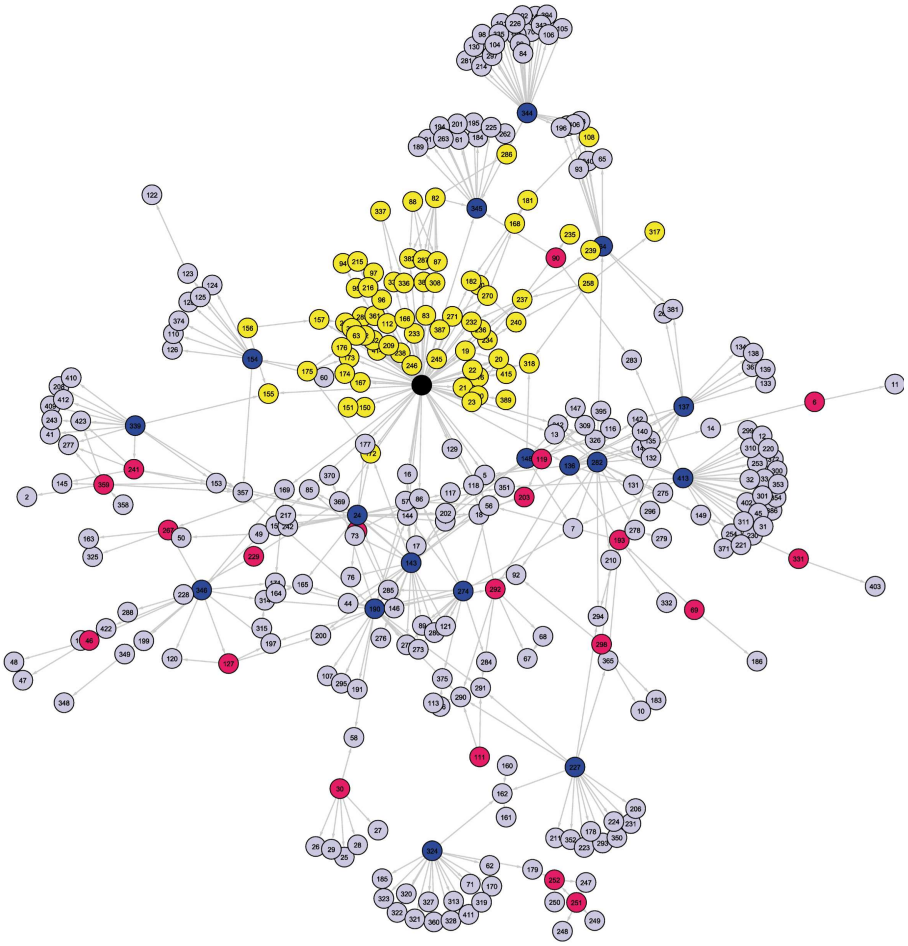

FIG. 1. Ajustement du modèle à blocs stochastiques avec $Q = 5$ classes au réseau de régulation d'*Escherichia coli* (source : [5]). Chaque niveau de gris correspond à une classe de nœuds.

des gènes d'un individu, soit de l'ordre de 22 000 variables pour l'être humain. Le nombre d'échantillons disponibles dépasse rarement la centaine. Les méthodes d'analyse différentielle (décrites précédemment) permettent de sélectionner les gènes prédisant au mieux une réponse d'intérêt, telle l'appartenance à un sous-ensemble de l'échantillon (individu sain/malade). L'inférence de réseau se situe naturellement dans la continuité de l'analyse différentielle, tâchant de déterminer parmi les gènes les plus explicatifs lesquels interagissent et régulent éventuellement les autres.

Un modèle couramment adopté en statistique et qui connaît un large succès dans les applications à l'inférence de réseau est celui des modèles graphiques gaussiens [6] : on décrit naturellement une expérience de puce à ADN à l'aide d'un vecteur aléatoire gaussien X de taille p dont la structure de corrélation est expliquée par la matrice $\mathbf{\Sigma} = (\Sigma_{ij})_{i,j\in\mathcal{P}}$ de variance-covariance. Chaque entrée du

vecteur X correspond à un gène. À partir de n puces notées $\mathbf{X} = (X^1, \ldots, X^n)$, l'estimation de la matrice de covariance doit permettre d'inférer la structure de dépendance entre les gènes, pour ainsi reconstruire un réseau. Une première idée pour reconstruire les interactions du réseau serait d'utiliser directement la covariance empirique entre les variables, se fondant sur le fait que[5] $X_i \perp\!\!\!\perp X_j \Leftrightarrow \Sigma_{ij} = 0$ dans le cas gaussien. Pourtant, de par la complexité des systèmes biologiques en jeu, les réseaux reconstruits uniquement sur cette base seraient très peu informatifs : les fortes corrélations entre gènes sont trop nombreuses et les effets de cascades aboutiraient à la détermination de réseaux « pleins », chaque gène étant lié aux autres. On utilise plutôt comme mesure d'interaction entre variables la notion de corrélation partielle, qui consiste à ne s'intéresser qu'aux interactions *directes* entre variables. On s'appuie alors sur le fait que les (in)dépendances *conditionnelles* sont décrites par l'inverse de la matrice de covariance empirique dans les modèles graphiques gaussiens[6] :

$$X_i \perp\!\!\!\perp X_j \mid \{X_{\mathcal{P}\setminus\{i,j\}}\} \Leftrightarrow \Sigma^{-1}_{ij} = 0.$$

Ainsi, la matrice $\mathbf{\Theta} = \mathbf{\Sigma}^{-1}$ décrit exactement le graphe de corrélations partielles ou de relations directes entre variables, et par là même entre les gènes. La log vraisemblance s'écrit en fonction de la covariance empirique $\mathbf{S}_n = \mathbf{X}'\mathbf{X}/n$ en vue de la détermination du paramètre d'intérêt, ici la matrice $\mathbf{\Theta}$:

$$\ell_{\mathbf{X}}(\mathbf{\Theta}) = \frac{n}{2}\log\det(\mathbf{\Theta}) - \frac{n}{2}\mathrm{Trace}(\mathbf{S}_n\mathbf{\Theta}) - \frac{n}{2}\log(2\pi). \tag{3}$$

Néanmoins, la maximisation de (3) pose problème : lorsque l'on dispose de suffisamment de données ($n > p$), on montre aisément que le maximum existe et est défini par $\mathbf{S}_n^{-1}$. Cependant, le graphe correspondant n'est que peu informatif puisque qu'aucune entrée de $\mathbf{S}_n^{-1}$ n'est nulle et toutes les arêtes du graphes sont actives. Par ailleurs, lorsque $n < p$, l'inversion n'est pas possible. Une approche extrêmement populaire issue de problèmes de régression consiste à *pénaliser* le problème pour contraindre $\mathbf{\Theta}$ à la forme voulue : le terme de pénalité vise non seulement i) à régulariser la solution ; ii) à induire de la parcimonie, c'est-à-dire des éléments exactement nuls dans le réseau reconstruit, afin de sélectionner les interactions les plus significatives. On peut même envisager d'introduire via cette pénalité un *a priori* sur la structure latente $\mathbf{Z}$ du réseau à reconstruire, issue de techniques présentées dans la section précédente ou de connaissance biologique. L'estimateur régularisé peut donc s'écrire comme le problème d'optimisation

$$\arg\min_{\mathbf{\Theta}} -\ell_{\mathbf{X}}(\mathbf{\Theta}) + \lambda \cdot \mathrm{pen}\,(\mathbf{\Theta}, \mathbf{Z})\,,$$

où le paramètre $\lambda > 0$ permet de doser le compromis en le terme de vraisemblance aux données et la forme attendue du réseau.

À titre d'exemple nous avons réalisé à l'aide des outils décrits dans [1] une étude sur un lot de puces à ADN concernant des patientes atteintes de cancer du sein. Certaines d'entre elles avaient une très bonne réponse à la chimiothérapie, d'autres

[5] Le symbole $\perp\!\!\!\perp$ est utilisé pour dénoter l'indépendance de deux variables aléatoires.

[6] La notation $X_i \perp\!\!\!\perp X_j \mid \cdot$ est utilisée pour l'indépendance conditionnelle, et l'ensemble $\mathcal{P}\setminus\{i,j\}$ correspond à toutes les variables de $\mathcal{P}$ sauf i et j. Ainsi $X_i \perp\!\!\!\perp X_j|\{X_{\mathcal{P}\setminus\{i,j\}}\}$ signifie que l'expression des gènes i et j est indépendante lorsque l'expression de tous les autres gènes est fixée.

ayant encore des résidus tumoraux après traitement. L'inférence de réseau, réalisée sur les puces correspondant à ces deux sous-ensembles, a permis d'identifier des interactions qui différaient entre les deux types de patientes (cf figure 2). Une étude poussée auprès des biologistes pourrait idéalement permettre de mettre à jour la défaillance de certains mécanismes de transcription chez les malades.

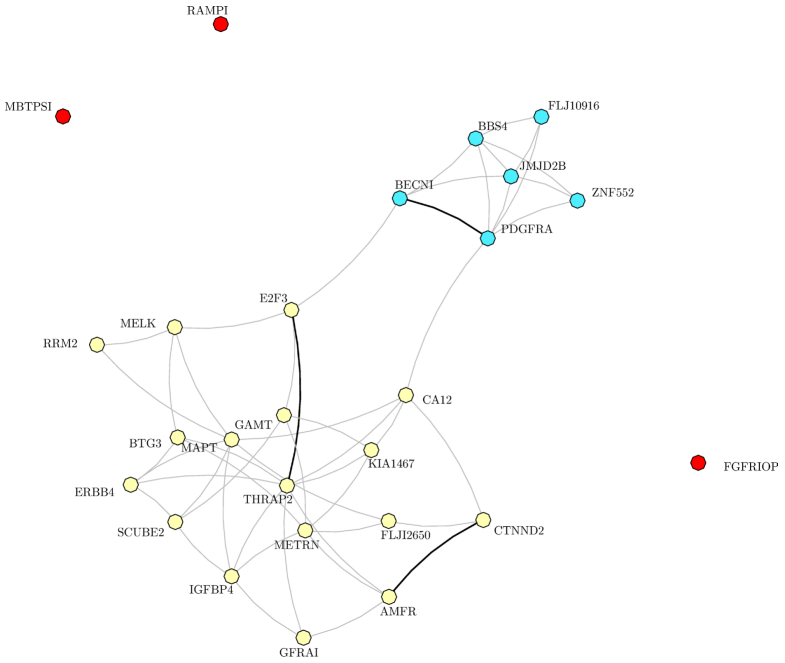

FIG. 2. Comparaison des réseaux reconstruits pour les données cancer du sein selon la réponse au traitement. Les traits épais indiquent les arêtes différant d'une série d'une patientes à l'autre.

4. Références

[1] C. AMBROISE ; J. CHIQUET ; C. MATIAS. Inferring sparse Gaussian graphical models with latent structure, *Electronic Journal of Statistics*, 3 : 205–238, 2009.

[2] A. BARABÁSI ; R. ALBERT. Emergence of scaling in random networks *Science* 286 (5439) : 509-512, 1999.

[3] J.-J. DAUDIN, F. PICARD, S. ROBIN, Mixture model for random graphs, *Statistics and Computing*, 2008.

[4] P. LATOUCHE, *Modèles de graphes aléatoires à structure cachée pour l'analyse des réseaux*, mémoire de thèse, 2011.

[5] Groupe de travail SSB – Statistics for Systems Biology, *page web du logiciel MixNet*, `http://stat.genopole.cnrs.fr/logiciels/mixnet`.

[6] J. WHITTAKER, *Graphical Models in applied multivariate Statistics*, 1996.

Le comité de rédaction remercie chaleureusement Patricia Reynaud-Bouret qui a coordonné ce dossier pour la *Gazette*.