2

Problématiques statistiques à l'heure de la post-génomique

Pierre Neuvial (2003), UC Berkeley, Department of Statistics,
et Pierre-Yves Bourguignon (2002), CEA/Institut de génomique,

Il est possible de caractériser et quantifier la vie cellulaire. Les enjeux sont désormais l'intégration et le traitement de ces données et le dépassement du 'fléau de la dimension' qu'elles posent : trop de variables et pas assez d'individus... Un défi théorique et pratique nouveau et lourd de conséquences pour les statistiques, tant un cadre de modélisation cohérent permettrait de poursuivre l'étude du vivant et d'en dériver des applications biomédicales, industrielles ou environnementales utiles.

L'élucidation, en 1954, de la structure en double hélice de l'ADN et du lien entre sa structure et sa fonction par MM. Watson et Crick, a ouvert la voie à l'investigation systématique des mécanismes moléculaires à l'œuvre dans le vivant. Ininterrompue à ce jour, l'accumulation de connaissances sur les entités chimiques mises en jeu et leurs interactions a permis d'élaborer progressivement une vision globale des flux d'information sous-tendant la vie, mais aussi d'enrichir le spectre de technologies à la disposition des biologistes afin d'observer ces mécanismes.

Aujourd'hui, il est possible de quantifier de nombreux aspects de la vie d'une cellule ou d'un organisme, et les enjeux de la recherche se situent à présent dans l'intégration de ces données au sein de cadres de modélisation cohérents qui permettent non seulement de poursuivre l'étude du vivant pour elle-même, mais aussi d'en dériver des applications biomédicales, industrielles, ou encore à la remédiation des problèmes environnementaux.

Un certain nombre d'initiatives récentes esquissent les contours des biotechnologies à venir, soulevant des questions dont les réponses relèvent aussi, en partie, du statisticien.

La diversité, caractéristique du vivant

Si l'ensemble des chromosomes portés par les être vivants à la surface de la Terre sont composés à partir des mêmes molécules, et si les séquences codantes des gènes inscrits sur ces chromosomes sont quasi invariablement traduites en protéines selon le même code génétique, les séquences de ces chromosomes n'en restent pas moins marquées par une diversité dont on ne peut que sous-estimer l'ampleur compte-tenu du relativement faible nombre de génomes séquencés.

Résultat de la combinaison des milliards d'années durant de l'infidélité inhérente aux processus de réplication de l'ADN aux processus de sélection naturelle, la diversité actuelle des séquences

génomiques sous-tend la diversité des espèces, mais également la diversité intra-spécifique. Elle affecte aussi bien la nature des protéines, que la dynamique de leur production par la cellule en réponse aux stimuli reçus de l'environnement.

En complément du séquençage et du typage des génomes, diverses technologies permettent aujourd'hui aux biologistes d'observer diverses facettes de la dynamique moléculaire dont une cellule est le siège, offrant l'opportunité de mettre à jour l'impact de chacune de leurs multiples variations individuelles sur le fonctionnement d'un organisme. Cette entreprise requiert des approches supervisées d'une part, et de combiner les résultats recueillis par les diverses méthodes expérimentales permettant de caractériser ces dynamiques d'autre part.

Elle se heurte néanmoins à un obstacle inhérent : tant de facteurs varient entre les classes de l'ensemble d'apprentissage que de nombreux artefacts apparaissent significatifs, masquant les éléments effectivement déterminants. Si la diversité engendrée par l'évolution constitue un défi pour les approches analytiques en biologie et en médecine, comme nous le verrons ci-après dans le cas de la cancérologie, elle est en revanche une fantastique opportunité pour la biologie synthétique.

Recherche en cancérologie, maladie des gènes

Le cancer est un enjeu majeur de santé publique : en 2007, plus de 12 millions de nouveaux cas ont été diagnostiqués dans le monde, et environ 7,6 millions de personnes sont décédées des suites d'un cancer. Cette maladie devrait être désignée par un pluriel, car les cancers constituent un ensemble hétérogène de maladies, qui se manifestent par une prolifération incontrôlée de cellules anormales, qui accumulent des modifications du génome les rendant capables d'envahir d'autres tissus (Hanahan et Weinberg, 2000).

La transformation d'une cellule normale en une cellule tumorale passe par l'altération des gènes qui contrôlent typiquement la croissance, la différenciation cellulaire, et la maintenance de l'ADN. Ces altérations du génome peuvent avoir lieu à différentes échelles (gains ou pertes de chromosomes partiels voire entiers, mutations affectant une seule lettre de la séquence d'ADN, variation de la longueur d'une séquence répétée), et peuvent aussi bien résulter de la seule dérive de l'ADN propre à la prolifération tumorale, qu'y contribuer plus ou moins directement.

La diversité intrinsèque aux évolutions tumorales fait de chaque patient le porteur d'une maladie différente, et constitue le défi majeur de la recherche en cancérologie dans ses entreprises de découvertes des causes du développement des cancers, de mise au point de cibles thérapeutiques, et d'adaptation des traitements aux spécificités de chaque patient.

La nécessité de comprendre et de caractériser les altérations des gènes dans les cancers a stimulé le développement de nouveaux outils de biologie moléculaire comme les puces à ADN et le séquençage, qui permettent d'interroger un très grand nombre de variables à différents niveaux d'information biologique dans la cellule : génotypage, cartographie des variations de nombre de copies d'ADN, niveau d'activité (ou expression) des gènes et des protéines.

Nouveaux types de données, nouveaux enjeux statistiques

Bien que les questions statistiques posées par la recherche en cancérologie puissent paraître classiques au premier abord, la dimension des données disponibles et la complexité des relations entre les variables nécessitent bien souvent le développement de méthodes statistiques originales, voire de nouvelles théories.

" La diversité intrinsèque aux évolutions tumorales fait de chaque patient le porteur d'une maladie différente et constitue le défi majeur : (...) plusieurs dizaines de milliers de gènes, pour quelques dizaines d'individus "

Prenons l'exemple de la recherche de marqueurs biologiques pronostics dans les cancers du sein, c'est-à-dire d'indicateurs permettant de prédire si une patiente opérée pour un cancer du sein va faire une rechute en l'absence de traitement anti-cancer suivant l'opération. Aujourd'hui, sur dix femmes opérées pour un cancer du sein, neuf reçoivent un traitement par chimiothérapie car leur risque de rechute est élevé d'après les indicateurs cliniques classiques dont on dispose. Ceux-ci sont fondés sur des propriétés morphologiques des cellules tumorales observables au microscope, comme le grade de la tumeur, mais manquent de spécificité : on sait que dans 70 % des cas, la chirurgie seule (non suivie de chimiothérapie) suffirait à éviter la rechute, mais on ne sait pas quels sont ces cas. La découverte de marqueurs pronostics plus pertinents devrait permettre d'éviter des traitements inutiles et très lourds en effets secondaires.

Les données à disposition du statisticien sont généralement les niveaux d'expression de plusieurs dizaines de milliers de gènes, pour quelques dizaines d'individus (quelques centaines au mieux) dont on sait qu'ils ont rechuté ou non après le traitement. Il s'agit donc d'un problème de classification : au vu de cet ensemble de données à réponse connue, peut-on élaborer une règle de décision permettant ensuite d'évaluer le risque de rechute d'un individu dont on connaîtrait seulement le niveau d'expression des gènes ?

Cependant, les méthodes de classification usuelles comme l'analyse discriminante ont été développées pour le cas où le nombre d'observations est supérieur au nombre de variables ; dans le cas des données post-génomiques, où le nombre de variables dépasse de plusieurs ordres de grandeurs le nombre d'observations, il peut exister une infinité de façons de combiner les variables pour séparer les deux groupes d'observations. Il s'agit d'une des manifestations du « fléau de la dimension », qui apparaît aussi dans d'autres domaines d'application des statistiques, notamment ceux liés à l'analyse d'image (imagerie médicale, astronomie).

De nouvelles méthodes ont donc été développées pour répondre aux problèmes de « classification en grande dimension » : sélection de variables avant ou pendant la classification, ajout de contraintes sur la forme de l'ensemble des prédicteurs, méthodes dédiées pour l'estimation de l'erreur. L'étude des propriétés théoriques de ces méthodes constitue désormais un nouveau pan de la statistique appliquée.

Une autre caractéristique importante des données post-génomiques est la forte dépendance entre les variables : on sait que les gènes d'une cellule agissent de façon coordonnée, en réseaux de régulation. Les méthodes statistiques classiques (notamment les méthodes de classification) font généralement l'hypothèse que les variables observées sont indépendantes. Un champ de recherche actif concerne d'une part la cartographie à grande échelle des relations entre ces variables (inférence de réseaux de régulation) et d'autre part l'intégration de l'information biologique disponible sur ces relations dans les modèles statistiques.

Vers une médecine personnalisée ?

Les études menées depuis une dizaine d'années à partir de données de puces à ADN ont permis d'affiner la classification moléculaire des cancers ainsi que la compréhension des mécanismes de leur développement, et également d'identifier de nombreux marqueurs biologiques pronostics candidats dans différents types de cancers.

La validation par des essais cliniques indépendants de ces candidats reste une étape coûteuse et longue, ne serait-ce que du fait du recul nécessaire pour constater l'absence de rechute. Cependant, quelques tests diagnostics basés sur les résultats de ces études sont déjà commercialisés : par exemple, le test Mammaprint développé par

" La complexité des relations entre les variables nécessite le développement de méthodes statistiques originales, voire de nouvelles théories "

la société néerlandaise Agendia est utilisé aux États-Unis et dans plusieurs pays européens pour évaluer le risque de rechute dans les cancers du sein : les résultats du test entrent en compte dans la décision du médecin de prescrire une chimiothérapie à la patiente.

Outre les marqueurs pronostics, qui visent donc à déterminer qui traiter, des marqueurs dits de prédiction ont également été développés, qui ont pour objectif de déterminer comment traiter, en prédisant la réponse du patient à un traitement ciblé. Ainsi un traitement à l'aide d'Herceptine sera efficace chez les patientes atteintes de cancers du sein dont le gène ERBB2 est amplifié (c'est-à-dire présent en un grand nombre de copies dans le génome).

Ces exemples illustrent l'évolution actuelle des traitements en cancérologie vers une prise en charge de plus en plus spécifique des caractéristiques du patient.

Cette tendance est poussée à l'extrême par des entreprises comme 23andMe qui proposent pour quelques centaines de dollars d'évaluer, à l'aide d'un génotypage à grande échelle, la prédisposition d'un individu à des maladies génétiques, au diabète, à certains cancers, mais aussi l'intolérance au lactose, à l'alcool...

D'une part, ces évaluations sont peu fiables du fait de notre connaissance incomplète de l'extrême complexité des facteurs de risque de la plupart des « traits » testés. D'autre part, la valeur ajoutée de ces tests du point de vue clinique est loin d'être démontrée car les stratégies de prévention actuelles (arrêter de fumer, avoir une alimentation saine et une activité physique régulière) ne sont pas spécifiquement adaptées à tel ou tel trait.

Faire du neuf dans le vivant, la biologie synthétique

Au cours de son investigation du vivant, la biologie a élucidé les capacités biochimiques d'un nombre conséquent de protéines ainsi qu'un certain nombre de mécanismes régulateurs de l'expression des gènes. Bien qu'issus d'organismes divers, nombre de ces mécanismes peuvent être clonés entre espèces, et constituent autant de briques élémentaires que l'on peut chercher à réassembler au sein d'un même organisme à des fins utilitaristes (Marlière, 2008). Les protéines présentent en effet cette vertu de catalyser un large spectre de réactions chimiques dans des

conditions domestiques, la catalyse purement chimique des mêmes réactions nécessitant souvent des apports énergétiques considérables et le recours à des composés dangereux.

La biologie synthétique ambitionne de systématiser la conception d'organismes vivants « reprogrammés » capables de réaliser une fonction d'intérêt, et entend ne pas se restreindre aux catalogues des fonctions naturelles : la prestation de synthèse d'un gène de séquence arbitraire coûte aujourd'hui 1 000 € ; pour le même prix, on peut également recevoir plusieurs millions de variations du même gène. Il en découle qu'il faut s'attendre à une expansion rapide du corpus génomique du vivant, que seule freine pour l'instant notre incapacité à rédiger un texte génomique à la fois nouveau et fonctionnel.

Il en résulte une double révolution pour la biotechnologie.

D'une part, de nombreux industriels sont d'ores et déjà convaincus que la biodiversité présente à la surface de la Terre ne suffira pas à satisfaire les besoins actuels et futurs d'une industrie plus respectueuse de l'environnement. Concevoir de nouveaux gènes que l'évolution naturelle des organismes n'aurait pu engendrer constitue donc une opportunité de mieux satisfaire les besoins.

D'autre part, on se doit de reconnaître un danger fondamental : celui de la pollution génétique (Rifkin, 1998). En effet, équiper des organismes de protéines exceptionnellement efficaces pour réaliser telle ou telle réaction augmente d'autant

le risque de voir cette fonction détournée par l'organisme, lui conférant ainsi une capacité accrue de prolifération, voire de pathogénicité.

Un rôle clef pour l'évolution

Les réponses proposées à ce jour pour répondre à ce double challenge de concevoir *de novo* des protéines à la fois utiles et domestiques accordent un rôle prépondérant aux modélisations statistiques et stochastiques.

Contrairement à ce que l'on peut lire sur le sujet, il n'est pas nécessaire de tout spécifier lorsque l'on reprogramme un organisme. En l'absence d'une compréhension suffisante du rôle fonctionnel de chaque nucléotide d'un génome, cette approche présente d'ailleurs peu de chances de succès : les fonctions que l'on implante ainsi ne résistent pas plus de quelques générations à l'évolution spontanée de l'organisme.

Une alternative consiste à induire des déviations radicales dans des séquences connues afin de réorienter l'évolution sur un chemin différent, et de laisser les organismes ainsi modifiés évoluer dans des conditions contrôlées sur des périodes suffisamment longues. Le bénéfice est double : la fonction implantée est évolutivement stable, et elle a bénéficié de l'optimisation locale permise par la sélection opérant pendant la phase d'évolution contrôlée.

Dans ce processus, le rôle du statisticien est primordial pour caractériser les aspects des protéines conservés par l'évolution naturelle, et donc pour cibler les mutations à même d'induire de telles déviations. Ces catalyseurs biochimiques que l'évolution n'aurait probablement jamais engendrés seule doivent être confinés au mieux afin qu'ils ne prolifèrent pas dans les organismes vivants naturels.

Au-delà du confinement physique, diverses stratégies de verrouillage de l'information génétique sont à l'étude aujourd'hui, et s'appuient également sur la possibilité de modifier radicalement, mais imparfaitement, un organisme, et de le laisser par la suite s'adapter au cours d'une convalescence adaptative. Le spectre des solutions proposées inclut la dépendance de l'organisme à un nutriment non-naturel, la modification du code génétique de l'organisme (grâce auquel un gène de l'organisme modifié n'a aucun sens pour un organisme naturel), la modification des monomères (nucléotides, acides aminés) constituant les macromolécules biologiques, etc...

La production d'organismes modifiés radicalement et destinés à servir de châssis dans lesquels implanter des fonctions d'intérêt industriel sera à moyen terme l'objet d'un véritable marché, pour lequel des normes seront nécessaires. Il faudra alors que les États se dotent de réglementations fondées sur des modèles de dynamique des populations, et destinées à contrôler la probabilité qu'un gène prolifère dans la nature compte-tenu du nombre et du type de verrous physiques, nutritionnels, ou informationnels auxquels est soumis le chromosome qui le porte, ainsi que de la taille de population permise par la mise en œuvre industrielle de l'organisme. ■

Références

D. Hanahan and R. A. Weinberg, 'The hallmarks of cancer', *Cell*, 100(1) :57-70, Jan 2000.

D. J. Hunter, M. J. Khoury, and J. M. Drazen, 'Letting the Genome out of the Bottle - Will We Get Our Wish?', *New England Journal of Medicine*, 358(2) :105-107, January 2008.

P. Marlière, 'Pourquoi et comment faire des formes de vie nouvelles ?', *Conférence de l'Université de tous les savoirs*, 2008, <http ://www.canal-u.tv/producteurs/universite_de_tous_les_savoirs/dossier_programmes/les_conferences_de_l_annee_2008/qu_est_ce_que_la_vie_ou_en_est_on_de_la_connaissance_du_genome/pourquoi_et_comment_faire_des_formes_de_vie_nouvelles_philippe_marliere>/

J. Rifkin, *The Biotech Century : Harnessing the Gene and Remaking the World*, Published by J P Tarcher, 1998. ISBN 0-87477-909-X, <http ://en.wikipedia.org/wiki/Special :BookSources/087477909X>

C. L. Sawyers, 'The cancer biomarker problem', *Nature*, 452(7187) :548-552, April 2008.

L. J. van 't Veer and R. Bernards, 'Enabling personalized cancer medicine through analysis of gene-expression patterns', *Nature*, 452(7187) :564-570, April 2008.