Big Data et connaissance biologique

Giuseppe Longo, Maël Montévil

Big Data et connaissance biologique[1]

Giuseppe Longo[2], Maël Montévil[3]

Résumé

Certains auteurs affirment que l'analyse des grandes bases de données pourrait remplacer la méthode scientifique. A contrario, nous argumentons que la bonne manière de faire fructifier ces nouveautés techniques est de les encadrer théoriquement. En biologie, en particulier, il nous semble urgent de développer une théorie des organismes.

1. Introduction

La biologie est un domaine où la variation a un rôle théorique fondamental. La variation biologique est profonde, qualitative, et nous avons défendu ailleurs l'idée qu'elle nécessite une épistémologie propre, notamment car la variation biologique fonde l'historicité et la contextualité du vivant [LON 14, MON16a].

Face à cette diversité du vivant et face à l'hétérogénéité interne des organismes, l'esprit humain semble parfois quelque peu démuni. La possibilité contemporaine de développer d'immenses bases de données numériques, de manière collaborative,

---

[1]G. Longo & M. Montévil. « Big Data et connaissance biologique ». In : *Sciences de la vie, sciences de l'information.* Sous la dir. de T. Gaudin, D. Lacroix, M.-C. Maurel et al. Paris : ISTE-Editions., 2017.

[2] Centre Cavaillès, République des Savoirs, CNRS USR3608, Collège de France et École Normale Supérieure, Paris, France et Department of Integrative Physiology and Pathobiology, Tufts University School of Medicine, Boston, MA USA.

[3] Laboratoire "Matière et Systèmes Complexes" (MSC), UMR 7057 CNRS, Université Paris 7 Diderot, 75205 Paris Cedex 13, France et Institut d'Histoire et de Philosophie des Sciences et des Techniques (IHPST) - UMR 8590.

semble alors une opportunité majeure. Mais cette opportunité n'est pas sans périls et l'analyse dénuée de sens biologique n'est pas le moindre de ces périls. Dans certains cas, les sciences biologiques sont armées pour maîtriser ces bases de données croissantes. L'analyse phylogénétique, par exemple, s'appuie sur le cadre conceptuel de la théorie de l'évolution, ce qui lui permet d'encadrer la production de connaissances à partir des données, et ceci avec des structures mathématiques non-triviales. Par contraste, il n'existe pas de théorie bien établie pour comprendre les organismes et leurs fonctionnements. Malgré des décennies d'usages informels, la notion de programme génétique n'a jamais acquis de substance théorique réelle. Cette tradition conduit néanmoins à une priorité causale assignée au niveau moléculaire, priorité qui se matérialise par la nature des données obtenues par les techniques à haut débit. Par contraste la modélisation d'un organe tel que le cœur requiert la prise en compte simultanée de plusieurs niveaux d'organisation [NOB 06]. De même, certains physiciens insistent sur l'importance des dimensions physique dans la détermination des phénomènes biologiques alors qu'elles ne sont pas associées à des techniques à haut débit ; par exemple, le jeu des forces dans une dynamique morphogénétique n'est mesuré ni en génomique ni en protéomique.

Ces questions sont cruciales, car de manière générale les analyses statistiques se basent sur des hypothèses qui ont d'abord une origine théorique, certes parfois informelle voire implicite. La capacité des bases de données à contribuer à la compréhension des phénomènes dépend du regard théorique encadrant l'utilisation de ces données et leur conférant du sens, ainsi que de la pertinence de ces données par rapport à un cadre théorique. Bref, il y a toujours un choix, parfois considéré comme "évident", voire unique, des observables à mesurer, d'une métrique, de critères d'approximation numériques.

L'application des big data au cancer, par exemple, se fait dans un cadre théorique particulier, où le procès de carcinogenèse est conçu comme l'apparition de cellules cancéreuses par accumulation de mutation somatiques : « the story of cancer is a story of how the body's complex coding systems go awry through the creation of self-perpetuating errors in cellular replication and growth » [SHA 14]. Or ce point de vue théorique rencontre des difficultés conceptuelles et empiriques majeures qui se matérialisent notamment par des retombées médicales extrêmement limitées, malgré des investissements conséquents. L'un des avocats les plus influents de cette théorie de la carcinogenèse souligne que nous sommes à nouveau face à « une complexité infinie » devant ces phénomènes [WEI 14].

Plutôt que d'aborder la situation comme la manifestation d'un problème théorique, comme le font certains auteurs proposant des points de vue alternatifs sur

la nature de la carcinogenèse [SON 99, BAK 11, SON 16], les big data apparaissent parfois comme une solution permettant de soigner le cancer sans passer par une remise en cause théorique.

De manière plus générale, l'absence d'un cadre théorique pour les organismes rend particulièrement séduisante une certaine rhétorique allant au-delà voir contre l'utilisation raisonnée des données. Un mythe se construit autour de l'omnipotence et de l'autonomie de l'analyse des bases de donnée. Pendant une décennie, plusieurs textes à succès, y compris celui de Chris Anderson [AND 08], racontent que les chiffres parlent d'eux-mêmes : « We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot ... Correlation supersedes causation, and science can advance even without coherent models, unified theories … No semantic or causal analysis is required ». L'idée est alors que les « data miners » soient capable de détecter des corrélations et d'orienter la décision sans avoir à effectuer ces discussions théoriques. Il ne s'agit alors plus d'enrichir la méthode scientifique, « obsolète », mais bien de la remplacer et en particulier de se passer de théorie. Ce point de vue est associé au slogan suivant lequel plus la base de donnée est grande, plus il est aisé de trouver des relations sur la base desquels agir.

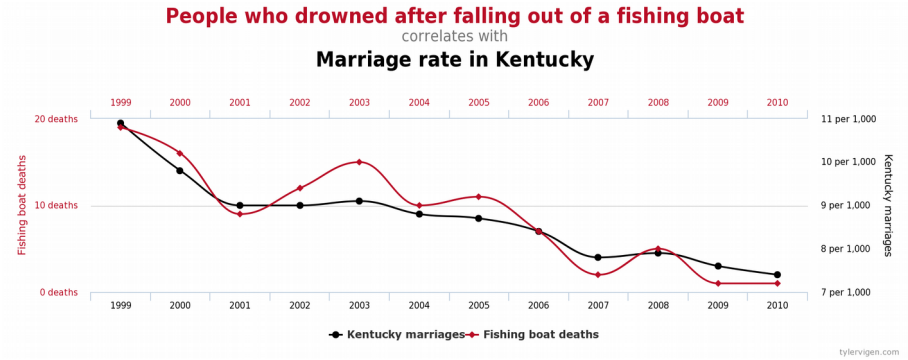
2. Grandes bases de données, prédiction et hasard

Les mathématiques permettent de démontrer les limites de ces méthodes purement algorithmiques, en montrant l'impossibilité de remplacer la quête scientifique du sens par du pur « data mining ». Des théorèmes à la croisée de la théorie ergodique et de la théorie de Ramsey, une théorie combinatoire des nombres née dans les années 20 et bien développée depuis, permettent de contredire cet usage des Big Data [CAL 16]. Les théorèmes "à la Ramsey", utilisés dans [CAL 16], montrent que pour toute corrélation entre nombres, il existe un nombre, *m* disons, tel que toute base de données ayant au moins *m* éléments contienne la corrélation demandée. Ce n'est donc qu'une question de taille, et il est possible de calculer un seuil au delà duquel on tombe *toujours* sur une "base de données" (un ensemble de nombres) qui contiendra une régularité avec les caractéristiques demandées. Autrement dit, précisez le critère de corrélation que vous souhaitez pour des paires, des triplets, etc., ainsi que le nombre minimal de fois que vous voulez l'observer, dans quel espace ou sur quelle durée, et que la manière dont vous départagez votre base de donnée (par exemple, en corrélant des valeurs proches, voire itérées … selon le critère préféré). Alors, les théorèmes mentionnés vous diront combien de données réunir pour y arriver. Plus précisément, une régularité dans un ensemble de nombres

peut être définie en fixant trois paramètres, voire plus ("arité" de la relation, cardinalité du seuil d'intérêt – combien vous en souhaitez, et la partition de la base de données ...) : à partir de ces paramètres, on peut alors calculer un nombre, *m*, tel que tout ensemble de nombres *A*, qui contienne au moins *m* éléments, satisfait la régularité demandée.

Il faut bien observer que *A* est un ensemble quelconque, il doit seulement être "assez grand", en fait énorme, car *m* est très fortement croissant en fonction des paramètres donnés. Mais, étant arbitraire, *A* peut être engendré par des … lancements de dés, des mesures du spin-up/spin-down d'un électron, un phénomène quantique aléatoire ou d'une quelconque nature, physique, biologique … Plus c'est grand, mieux c'est, nous disent les propagandistes a-critiques des Big Data ? Ce nombre *m* est trop grand pour être rencontré dans notre Univers pour une corrélation entre un nombre suffisant d'éléments ? Il faudrait alors donner des seuils *en dessous* desquels on puisse appliquer la technique brute du « data mining », sans tomber dans les limites posées par ces théorèmes. En tout cas, ces résultats nous disent que *tout A* suffisamment grand contient la corrélation arbitraire pré-donnée : si l'on se contente qu'elle apparaisse dans un pourcentage plus bas que 100% des ensembles, par exemple de sorte à avoir "seulement" un pourcentage raisonnablement haut d'ensembles "aléatoires", donc de corrélations certainement « spurious », on obtiendra des *m* atteignables par nos bases de données. Bref, cet hasard dans les grandes quantités de nombres n'est point rare. Expliquons-nous.

Un ensemble fini de nombres est, grossièrement, dit « aléatoire », lorsqu'il ne peut pas être engendré par un programme plus petit que son nombre d'éléments. Or, le pourcentage des ensembles de nombres aléatoires en ce sens faible tend vers 100 % (mesure 1, pour être plus précis) quand leur cardinalité croît vers l'infini. Or, l'infini est grand, même pour les "data miners" les plus riches en données, mais dès qu'on a à faire à des ensembles de nombres qui s'expriment avec 2000 bits, par exemple – ce qui n'est pas hors mesure, on frôle le 80 % d'ensembles incompressibles, [CAL 16]. Bonne chance donc pour faire un quelconque usage en termes de prédiction ou d'action avec des données qui peuvent dériver du hasard. Car, dans tout ces cas où le hasard domine, il est hors de question que les régularités trouvées par des programmes astucieux d'explorations des données puissent aider à prédire, voire à agir, car, justement, elles sont le fruit du hasard, et elles peuvent donc ne pas se reproduire, dans le temps, dans l'espace, ni dériver d'aucune relation causale …. C'est ainsi, que, à tout hasard, on tombe dans des corrélations comme celle représentée ci-dessous (extrait de "Spurious correlations" http://www.tylervigen.com/spurious-correlations, Nov., 2015") :



Dans un objectif d'action, nous avons immédiatement écrit au gouverneur du Kentucky, pour qu'il interdise les mariages ….

Les résultats cités ici sont techniques : ils appartiennent à la théorie combinatoire des nombres et à la théorie des algorithmes. Les défenseurs des « Big Data sans théories » et des algorithmes de data mining sans analyses du sens, ignorent, par principe, les cadres théoriques. Or, la théorie combinatoire des nombres et la théorie des algorithmes démontrent leurs propres limites dans les possibilités de calcul et de prédiction, par ces « résultats négatifs » qui en sont à l'origine et sont propres à la connaissance scientifique. Plus particulièrement, des variantes des résultats de la théorie de Ramsey se situent près de l'espace difficile de ce qui est calculable, mais dont on ne peut pas démontrer la calculabilité dans la théorie formelle des nombres [LON 11]. Une fois saisi l'importance des limites du "tout algorithmique", du "tout calculable", on peut alors utiliser au mieux ces immenses quantités de donnée que l'informatique rend disponible, une grande chance pour la science, dans tous les domaines, dont la biologie. Une fois clarifiées les hypothèses qui font choisir certains observables et pas d'autres et le choix de mesures adéquates aux objectifs de connaissance que l'on se donne, les informations numériques peuvent aider à la conjecture, à la corroboration d'une théorie ou à son esquisse, voire à de nouvelles compréhensions. Qu'elle précède ou qu'elle soit impulsée par l'analyse des données, il nous semble cependant urgent et nécessaire de développer la réflexion théorique pour la compréhension des organismes. Dans ce contexte, nous sommes engagés dans un effort collaboratif et interdisciplinaire dont les derniers résultats forment un numéro spécial de Progress in Biophysics and Molecular Biology : From the century of the genome to the century of the organism: New theoretical approaches, [SOT 16].

7. Bibliographie

La plupart des articles des auteurs de ce chapitre peuvent être téléchargés depuis leurs pages web.

[AND 08] ANDERSON, C. « The end of theory: The data deluge makes the scientific method obsolete. » *WIRED*. 2008.

[BAK 11] BAKER, S. G. 2011, TOFT better explains experimental results in cancer research than SMT. Bioessays, 33: 919–921. doi:10.1002/bies.201100124

[BAI 06] BAILLY, F., LONGO, G. « Mathématiques et sciences de la nature. La singularité physique du vivant. » Hermann, Paris, 2006.

[CAL 16] CALUDE, C, LONGO, G. « The Deluge of Spurious Correlations in Big Data », to appear in *Foundations of Science*, 2016.

[LON 11] LONGO, G. « Reflections on Concrete Incompleteness », in *Philosophia Mathematica*, 19(3): 255-280, 2011.

[LON 14] LONGO, G., MONTÉVIL, M. *Perspectives on Organisms: Biological Time, Symmetries and Singularities.* Springer, Berlin. 2014.

[MON 16] MONTÉVIL, M., MOSSIO, M., POCHEVILLE, A., LONGO, G.. « Theoretical principles for biology: Variation », *Progress in Biophysics and Molecular Biology,* Available online 13 August 2016.

[NOB 06] NOBLE, D. *The Music of Life: Biology beyond the Genome.* Oxford University Press, Oxford. 2006.

[SHA 14] SHAIKH AR, BUTTE AJ, SCHULLY SD, DALTON WS, KHOURY MJ, HESSE BW « Collaborative Biomedicine in the Age of Big Data: The Case of Cancer » *J Med Internet Res* ; 16(4):e101 2014.

[SON 99] SONNENSCHEIN, C., SOTO, A.M. *The Society of Cells: Cancer and Control of Cell Proliferation.* Springer Verlag, New York. 1999.

[SON 16] SONNENSCHEIN, C., SOTO, A.M., « Carcinogenesis explained within the context of a theory of organisms ». *Progress in Biophysics and Molecular Biology*2016.

[SOT 16] SOTO, A.M., LONGO, G. « Why do we need theories? ». *Progress in Biophysics and Molecular Biology*. 2016.

[WEI 14] Weinberg, RA, « Coming Full Circle—From Endless Complexity to Simplicity and Back Again », *Cell*, Volume 157, Issue 1, 27, Pages 267-271, ISSN 0092-8674. 2014.