

Leçons apprises de l'évaluation des méthodes d'inférence de réseau de gènes chez *Bacillus subtilis*

Léa GRIMA¹, Céline BROUARD^{1,†}, Simon DE GIVRY¹, Anne GOELZER², Élise MAIGNÉ¹, and Nathalie VIALANEIX¹

¹ Université de Toulouse, INRAE, UR MIAT, 31326, Castanet-Tolosan, France

² INRAE, UR1404, MaIAGE, Université Paris-Saclay, Jouy-en-Josas, France

{prenom.nom}@inrae.fr

† Presenting author.

De nombreuses approches statistiques [1, 2, 3, 4, 5, 6, 7] se proposent d'inférer des réseaux biologiques à partir de données d'observation ou d'intervention dans le but de modéliser les phénomènes de régulation au sein des cellules ou du tissu étudiés. Toutefois, les études permettant d'évaluer la qualité de ces méthodes d'un point de vue biologique sont rares et souvent basées soit sur des données simulées [8, 9], soit sur des mesures indirectes [10], soit sur des informations *a priori* partielles très générales (comme les réseaux PPI), dans une approche non-supervisée ou semi-supervisée [11, 12].

Dans cette présentation, nous présentons un travail d'évaluation des méthodes d'inférence basé sur la connaissance d'un réseau de régulation réel et exhaustif de la bactérie *Bacillus subtilis* [13]. Ce travail se fonde sur les données d'expression générées dans le cadre du projet européen de biologie des systèmes BaSysBio¹ et obtenues à partir de puces ADN haute résolution de type *Tiling Array* [14]. Ce jeu de données regroupe 269 individus (expériences) correspondant à 104 conditions de culture différentes conçues pour obtenir une couverture quasi-exhaustive d'expression du génome (seulement 4% des gènes n'ont été exprimés dans aucune condition).

Les résultats préliminaires ont permis la comparaison de méthodes basées sur la corrélation de Pearson, sur la corrélation partielle (package R **PCIT** [5]) et sur les forêts aléatoires (package R **GENIE3** [6]). Ils montrent que, sans information *a priori*, la précision des méthodes d'inférence à l'échelle du génome reste assez faible et que les méthodes les plus simples (seuillage de la corrélation) peuvent s'avérer aussi efficaces que des modèles plus complexes. Également, ils montrent que, si les prédictions des mécanismes individuels de régulation individuels entre gènes sont peu retrouvés (faible pertinence des arêtes), certaines voies de régulation liées à des mécanismes physiologiques particuliers (comme la sporulation) sont, au contraire, plutôt bien retrouvés. De manière plus précise, les clusters de gènes du réseau inféré (obtenus avec une méthode de classification de sommets de graphe, l'algorithme de Louvain [15]) contiennent des gènes régulés par les mêmes régulateurs ou les mêmes combinaisons de régulateurs.

References

- [1] J. Pearl and S. Russel, *Bayesian Networks*. Cambridge, Massachusetts, USA: Bradford Books (MIT Press), 2002.
- [2] N. Meinshausen and P. Bühlmann, "High dimensional graphs and variable selection with the Lasso," *Annals of Statistic*, vol. 34, no. 3, pp. 1436–1462, 2006.
- [3] J. Friedman, T. Hastie, and R. Tibshirani, "Sparse inverse covariance estimation with the graphical lasso," *Biostatistics*, vol. 9, no. 3, pp. 432–441, 2008.
- [4] P. Langfelder and S. Horvath, "WGCNA: an R package for weighted correlation network analysis," *BMC Bioinformatics*, vol. 9, p. 559, 2008.
- [5] N. S. Watson-Haigh, H. N. Kadarmideen, and A. Reverter, "PCIT: an R package for weighted gene co-expression networks based on partial correlation and information theory approaches," *Bioinformatics*, vol. 26, no. 3, pp. 411–413, 2010.

¹<http://basysbio.com/>

- [6] H.-T. Vân Anh, I. Alexandre, W. Louis, and G. Pierre, “Inferring regulatory networks from expression data using tree-based methods,” *PLoS ONE*, vol. 5, no. 9, p. e12776, 2010.
- [7] G. Monneret, F. Jaffrézic, A. Rau, T. Zerjat, and G. Nuel, “Identification of marginal causal relationships in gene networks from observational and interventional expression data,” *PLoS ONE*, vol. 12, no. 3, p. e0171142, 2017.
- [8] P. Mendes, W. Sha, and K. Ye, “Artificial gene networks for objective comparison of analysis algorithms,” *Bioinformatics*, vol. 19, no. Suppl_2, pp. ii122–ii129, 2003.
- [9] D. Marbach, J. C. Costello, R. Küffner, N. Vega, R. J. Prill, D. M. Camacho, K. R. Allison, the DREAM5 Consortium, M. Kellis, and G. Collins, James J. and Stolovitsky, “Wisdom of crowds for robust gene network inference,” *Nature Methods*, vol. 9, no. 8, pp. 796–804, 2012.
- [10] L. P. Diaz and P. Stumpf, Michael, “Gaining confidence in inferred networks,” *Scientific Reports*, vol. 12, 2022.
- [11] C. Brouard, M. Szafranski, and F. d’Alché Buc, “Semi-supervised penalized output kernel regression for link prediction,” in *Proceedings of the 28th International Conference on Machine Learning (ICML 2011)* (L. Getoor and T. Scheffer, eds.), 2011.
- [12] S. Chen and J. C. Mar, “Evaluating methods of inferring gene regulatory networks highlights their lack of performance for single cell gene expression data,” *BMC Bioinformatics*, vol. 19, p. 232, 2018.
- [13] J. P. Faria, R. Overbeek, R. C. Taylor, N. Conrad, V. Vonstein, A. Goelzer, V. Fromion, M. Rocha, I. Rocha, and C. S. Henry, “Reconstruction of the regulatory network for *Bacillus subtilis* and reconciliation with gene expression data,” *Frontiers in Microbiology*, vol. 7, p. 275, 2016.
- [14] P. Nicolas, U. Mäder, E. Dervyn, T. Rochat, A. Leduc, N. Pigeonneau, E. Bidnenko, E. Marchadier, M. Hoebeke, S. Aymerich, D. Becher, P. Bisicchia, E. Botella, O. Delumeau, G. Doherty, E. L. Denham, M. J. Fogg, V. Fromion, A. Goelzer, A. Hansen, E. Härtig, C. R. Harwood, G. Homuth, H. Jarmer, M. Jules, E. Klipp, L. Le Chat, F. Lecointe, P. Lewis, W. Liebermeister, A. March, R. A. Mars, P. Nannapaneni, D. Noone, S. Pohl, B. Rinn, F. Rügheimer, P. K. Sappa, F. Samson, M. Schaffer, B. Schwikowski, L. Steil, Stülke, T. Wiegert, K. M. Devine, A. J. Wilkinson, J. M. van Dijl, M. Hecker, U. Völker, P. Bessières, and P. Noirot, “Condition-dependent transcriptome reveals high-level regulatory architecture in *Bacillus subtilis*,” *Science*, vol. 335, no. 6072, pp. 1103–1106, 2012.
- [15] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, “Fast unfolding of communities in large networks,” *Journal of Statistical Mechanics: Theory and Experiment*, vol. P10008, pp. 1742–5468, 2008.