frontiers in GENETICS

EDITORIAL
published: 16 July 2014
doi: 10.3389/fgene.2014.00221

Quantitative assessment and validation of network inference methods in bioinformatics

Benjamin Haibe-Kains[1,2]* and Frank Emmert-Streib[3]*

[1] *Bioinformatics and Computational Genomics, Princess Margaret Cancer Centre, University Health Network, Toronto, ON, Canada*
[2] *Medical Biophysics Department, University of Toronto, Toronto, ON, Canada*
[3] *Computational Biology and Machine Learning Laboratory, Center for Cancer Research and Cell Biology, Queen's University Belfast, Belfast, UK*
**Correspondence: bhaibeka@uhnresearch.ca; v@bio-complexity.com*

Edited by:
Richard D. Emes, University of Nottingham, UK

Reviewed by:
Matt Loose, University of Nottingham, UK

Keywords: computational biology, network inference, computational genomics, personalized medicine, bioinformatics, network validation

The last years following the completion of the human genome project (Quackenbush, 2011) have given raise to major breakthroughs in the development of novel biotechnologies, such as next-generation sequencing, that sparked the generation of high-throughput "omics" data. The robustness and the cost-efficiency of these technologies increasing over time enabled the conduction of large screening experiments containing hundreds and even thousands of samples. As a consequence of these "big" biological and biomedical high-throughput datasets, advanced statistical methodology can now be employed requiring such large sample sizes.

This is one reason explaining the recent interest in methods that aim to infer biological networks. These methods offer the opportunity for better understanding the interactions between genomic features and the overall structure and behavior of the underlying networks. In order to foster this research direction we edited a Research Topic entitled "Quantitative Assessment and Validation of Network Inference Methods in Bioinformatics." This research topic was perceived as relevant and timely by the scientific community and we consequently received 15 contributions from research groups all over the world (Boucher and Jenna, 2013; Chun et al., 2013; de Matos Simoes et al., 2013; Lopes and Bontempi, 2013; Qian and Dougherty, 2013; Schrynemackers et al., 2013; Scott-Boyer et al., 2013; Staiger et al., 2013; Tran et al., 2013; Ho et al., 2014; Horn et al., 2014; Montojo et al., 2014; Olsen et al., 2014; Peng and Schork, 2014; Santra, 2014).

The topics addressed by these contributions can be broadly grouped into the following categories:

- Data integration (Boucher and Jenna, 2013; Chun et al., 2013; Scott-Boyer et al., 2013; Ho et al., 2014; Horn et al., 2014; Olsen et al., 2014; Santra, 2014)
- Network validation (de Matos Simoes et al., 2013; Lopes and Bontempi, 2013; Qian and Dougherty, 2013; Schrynemackers et al., 2013; Montojo et al., 2014; Olsen et al., 2014)
- Network inference (Lopes and Bontempi, 2013; Schrynemackers et al., 2013)
- Time series data (Lopes and Bontempi, 2013)
- Network interpretation (Boucher and Jenna, 2013; Chun et al., 2013; de Matos Simoes et al., 2013; Montojo et al., 2014; Scott-Boyer et al., 2013; Tran et al., 2013)
- Diagnostic applications (Staiger et al., 2013; Peng and Schork, 2014)
- Network modeling (Tran et al., 2013)

First of all, it is important to note that there is still no commonly accepted term to denote 'networks' that are inferred from gene expression data, which the vast majority of the contributed papers used for their inference. Indeed, depending on the context, these networks are called gene regulatory networks (de Matos Simoes et al., 2013; Lopes and Bontempi, 2013; Qian and Dougherty, 2013; Santra, 2014), molecular interaction networks (Horn et al., 2014; Olsen et al., 2014), gene co-expression networks (Scott-Boyer et al., 2013) or biological networks (Schrynemackers et al., 2013). We believe that this plurality denotes the diversity of usages and interpretations of such networks, while it may also reflect the lack of agreement due to the interdisciplinary nature of network inference in Bioinformatics. For the future it would be beneficial to find a common terminology for such networks, because this would certainly enhance the communicability within the community. At the moment, the term 'gene regulatory networks' seems to be the most frequent denotation in use, however, a thorough discussion of this important topic seems indispensable.

The two topics that attracted most interest in the submitted contributions are network validation and data integration. The former is a good reminder that the assessment of inferred networks is not trivial due to two major reasons. First, we still have only partial knowledge about gene regulatory networks even in organisms like *Saccharomyces cerevisiae* (yeast) or *E. coli*, which are considerably simpler than Human. Second, networks are structured objects that means we cannot only assess errors on the global scale for the whole network, but also on intermediate levels down to single interactions and any combination thereof, e.g., motifs or modules (Emmert-Streib and Altay, 2010). In addition, for labeled data enabling the usage of supervised learning methods further issues need to be addressed, as indicated and discussed in the review paper by Schrynemackers et al. (2013).

The integration of different datasets, either of the same or of different types, is certainly a topic that will gain even more attention in the future when more and new high-throughput technologies become available and the access to such datasets is simplified by a policy change of funding agencies making it

imperative for grant holders to provide free access to such data. It appears that Bayesian methods (Santra, 2014) provide a natural framework that is particularly suited for such an integration because of its flexibility and widespread acceptance as a fundamental statistical inference paradigm. However, other methods have also been proposed to tackle the challenge of heterogeneous data integration, such as the regression-based framework integrating priors extracted from the biomedical literature and other sources (Olsen et al., 2014). This provides opportunities for comparing novel methodological developments with well-established statistical approaches. We would like to emphasize that networks inferred from the integration of different datasets require a reassessment of their validation for similar reasons as for a supervised learning of gene regulatory networks (Schrynemackers et al., 2013).

For the future, we think that applications of inferred network, e.g., for diagnostic, predictive or therapeutic purposes in medicine will become very important for translational research because of their potential to provide a systems-approach, certainly required to understand complex disorders like cancer. However, until we reach this point more work is needed. For our Research Topic, two contributions have been submitted that are good examples for a better understanding of this problem. In Peng and Schork (2014) the authors found that network centrality measures, which are characterizing the importance of nodes within a gene network that has been constructed from the gene expression patterns, can be used to identify therapeutic targets. In contrast, in Staiger et al. (2013) the authors showed that current composite-feature classification methods considering a network structure, do not outperform simple single-genes classifiers in predicting outcome in breast cancer for prognostic purposes. It is interesting to note that the outcome of both studies allows opposing conclusions. Whereas the results in Peng and Schork (2014) can be seen as an encouragement for further studies employing network-based approaches, the results in Staiger et al. (2013) do not support this. However, by changing the perspective, the study by Staiger et al. (2013) suggests that we do not need to focus on single-gene studies because we can get similar results from network-based approaches. Now, the crucial question is which perspective should we chose? The choice of perspective actually depends on the use of the inferred networks, and therefore the goal of the study. On the one hand, if one is interested in building a predictive model, which does not need to be interpretable (often referred to as "black box" in the literature), then only performance of the inferred model matters; in this case scenario Staiger et al. (2013) showed that, for cancer prognosis, network-based approaches may not be relevant as they do not outperform simpler methods (singe genes). On the other hand, if one is more interested in the biological knowledge that could be extracted from statistical models, network-based approaches are extremely relevant as they are efficient ways to represent complex biological patterns while retaining good predictive ability.

Overall, we believe that, in a translational application, the underlying choice of perspective is of central importance. That means the utility of a network-based approach is expected to depend crucially on the biological question to which such a method should be applied to.

REFERENCES

Boucher, B., and Jenna, S. (2013). Genetic interaction networks: better understand to better predict. *Front. Genet.* 4:290. doi: 10.3389/fgene.2013.00290

Chun, H., Chen, M., Li, B., and Zhao, H. (2013). Joint conditional gaussian graphical models with multiple sources of genomic data. *Front. Genet.* 4:294. doi: 10.3389/fgene.2013.00294

de Matos Simoes, R., Dehmer, M., and Emmert-Streib, F. (2013). B-cell lymphoma gene regulatory networks: biological consistency among inference methods. *Front. Genet.* 4:281. doi: 10.3389/fgene.2013.00281

Emmert-Streib, F., and Altay, G. (2010). Local network-based measures to assess the inferability of different regulatory networks. *IET Syst. Biol.* 4, 277–288. doi: 10.1049/iet-syb.2010.0028

Ho, Y.-Y., Cope, L. M., and Parmigiani, G. (2014). Modular network construction using eQTL data: an analysis of computational costs and benefits. *Front. Genet.* 5:40. doi: 10.3389/fgene.2014.00040

Horn, F., Rittweger, M., Taubert, J., Lysenko, A., Rawlings, C., and Guthke, R. (2014). Interactive exploration of integrated biological datasets using context-sensitive workflows. *Front. Genet.* 5:21. doi: 10.3389/fgene.2014.00021

Lopes, M., and Bontempi, G. (2013). Experimental assessment of static and dynamic algorithms for gene regulation inference from time series expression data. *Front. Genet.* 4:303. doi: 10.3389/fgene.2013.00303

Montojo, J., Zuberi, K., Shao, Q., Bader, G., and Morris, Q. (2014). Network assessor: an automated method for quantitative assessment of a network's potential for gene function prediction. *Front. Genet.* 5:123. doi: 10.3389/fgene.2014.00123

Olsen, C., Bontempi, G., Emmert-Streib, F., Quackenbush, J., and Haibe-Kains, B. (2014). Relevance of different prior knowledge sources for inferring gene interaction networks. *Front. Genet.* 5:177. doi: 10.3389/fgene.2014.00177

Peng, Q., and Schork, N. (2014). Utility of network integrity methods in therapeutic target identification. *Front. Genet.* 5:12. doi: 10.3389/fgene.2014.00012

Qian, X., and Dougherty, E. (2013) Validation of gene regulatory network inference based on controllability. *Front. Genet.* 4:272. doi: 10.3389/fgene.2013.00272

Quackenbush, J. (2011). *The Human Genome: The Book of Essential Knowledge*. New York, NY: Imagine Publishing, Curiosity Guides.

Santra, T. (2014). A bayesian framework that integrates heterogeneous data for inferring gene regulatory networks. *Front. Bioeng. Biotechnol.* 2:13. doi: 10.3389/fbioe.2014.00013

Schrynemackers, M., Kueffner, R., and Geurts, P. (2013). On protocols and measures for the validation of supervised methods for the inference of biological networks. *Front. Genet.* 4:262. doi: 10.3389/fgene.2013.00262

Scott-Boyer, M.-P., Haibe-Kains, B., and Deschepper, C. F. (2013). Network statistics of genetically-driven gene co-expression modules in mouse crosses. *Front. Genet.* 4:291. doi: 10.3389/fgene.2013.00291

Staiger, C., Cadot, S., GyŽrffy, B., Wessels, L. F., and Klau, G. W. (2013). Current composite-feature classification methods do not outperform simple single-genes classifiers in breast cancer prognosis. *Front. Genet.* 4:289. doi: 10.3389/fgene.2013.00289

Tran, V., McCall, M. N., McMurray, H., and Almudevar, A. (2013). On the underlying assumptions of threshold boolean networks as a model for genetic regulatory network behavior. *Front. Genet.* 4:263. doi: 10.3389/fgene.2013.00263

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 17 June 2014; accepted: 26 June 2014; published online: 16 July 2014.
*Citation: Haibe-Kains B and Emmert-Streib F (2014) Quantitative assessment and validation of network inference methods in bioinformatics. Front. Genet. **5**:221. doi: 10.3389/fgene.2014.00221*
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics.

Copyright © 2014 Haibe-Kains and Emmert-Streib. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.