Journal of
Animal Breeding and Genetics

J. Anim. Breed. Genet. ISSN 0931-2668

EDITORIAL

What is a good (gene) network?

We are fashion victims. Networking is fashionable so we decided last year that it was time to start to work on networks. In the last years, biology has evolved to understand how the relationships between a large number of elements (genes, proteins, …) can influence the way a living organism functions. This question is well modelled by the use of biological networks that are a huge topic of interest in recent literature. As an example, the Leipzig WCGALP included several articles about biological networks (e.g. Tesson *et al.*; Jager *et al.*; Kadarmideen *et al.*; Reverter *et al.*). One of those was our work (Liaubet *et al.*) which, at the end of the talk, gave rise to THE question: 'What is a good network?' This innocent and apparently simple question has haunted us ever since. The answer is complex, much too complex for a short editorial, but we can draw some evidence from our experience.

When working with biological networks, we are dealing with many different underlying questions: gene networks, protein networks or when speaking about the kind of relationships that they model, transcriptomic networks, regulation networks and interaction networks. We may consider the particular case of a gene co-expression network based on high throughput transcriptomic data. Usually, in this field, very limited prior biological knowledge is available as well as a frustrating annotation level (usually, in that kind of experiment in livestock species, about half of the genes have no functional or ontological annotation). With that restricted background, the use of a network model can help to improve the knowledge about the way genes interact and to emphasize key genes implicated in a given process. In Leipzig, we admired Trudy Mackay's talk because the *Drosophila* model is so powerful.

However, network inference with that kind of data has to be handled with care. For example, our first attempt was to build co-expression networks of differential genes (genes whose expression varies according to a phenotype of interest) for a developmental trait, between species. The results were very disappointing because networks were unstable, had similar structure, whatever the kind of genes considered (differential or chosen at random), and no pertinent biological conclusion could be drawn. This was the perfect example of a bad network! In that first experiment, the main problem was the too low number (about five) of observations available for each species. However, it was the starting point to understand the key features needed to obtain a good network. Now note that the question 'What is a good network?' can be divided into two sub-questions: What is a good network for biologists? What is a good network for statisticians?

Statisticians like robustness. The number of observations used to define the network is never large enough. A simple simulation study can illustrate the fact that at least 20–30 observations are needed to accurately estimate a correlation coefficient, and even more to infer a medium size network (Schäfer and Strimmer, 2005a, *Bioinformatics* **21**, 754–764). Furthermore, methods designed to deal with 'small' sample sizes and a large number of variables are required: When using the common Graphical Gaussian Models, this can consist of regularization (Dobra *et al.*, 2004, *J. Multiv. Anal.* **90**, 196–212; Mainshausen and Bühlmann, 2006, *Ann. Stat.* **34**, 3), shrinkage (Schäfer and Strimmer, 2005b, *Stat. Appl. Genet. Mol. Biol.* **4**, 32), bootstrap (Schäfer and Strimmer, 2005a), etc. But once the strength of the relations between two genes has been inferred by the chosen method, the subsequent question is: What are the important interactions? No other processing could be performed (e.g., keeping the complete network with edges weighted by the partial correlations), but use of *ad hoc* or significant thresholds can be beneficial for the readability and further interpretability of the network.

Biologists like simple outputs to understand the data and recognize biological processes and molecular pathways. Moreover, they are on a quest for the 'Grail of causality', with the help of statistical models (Schadt *et al.*, 2005, *Nat. Genet.* **37**, 710–717; the Leipzig WCGALP papers by Rosa *et al.*; Tesson *et al.*; Mackay), or based on the properties of a gene network. In a good network, one should be able to emphasize the key genes in the biological process: For instance, hubs [nodes (genes) that are connected with a large number of genes] are straightforward candidates for being interesting genes but they can sometimes be disappointing because they are much

  doi:10.1111/j.1439-0388.2010.00916.x

too obvious or can hide other interesting features. Moreover, a good network could be divided into relevant groups of genes working together (Mao *et al.*, 2009, *BMC Bioinformatics* **10**, 34) because it is useful to stress out the macro-structure of networks by separating modules. Hence, there is a huge need for statistical methods designed for graph mining such as clustering (that isolates groups of highly inter-connected genes). Even more interesting is to relate modules to a small subset of biological functions. At that point, biologists would benefit from easy bioinformatic tools enabling the collection of additional information on interesting genes, such as mapping and annotation. With the list of key biological functions, finding what is expected is reassuring but should not be the final step of the analysis. Approaching the unknown is surely the most exciting part of untargeted approaches: In that field, using networks can provide indications to decipher the way a large group of genes work together and, by association with what is already known, to understand the role of each gene, even those that are not yet annotated or studied.

In conclusion, a standard approach to work with gene co-expression networks might entail the following: Using a very large body of observations to select several hundred interesting genes, a gene network could be inferred through a robust approach. Then, by clustering the genes from the network structure, we could obtain a small number of groups. In the ideal case, each group would be related to a single biologically relevant function. Such a conclusion can be obtained and would comfort us – and has done so – in our involvement in this fashionable research direction. Furthermore, from that conclusion, a few facts can be derived about a good network: It is a network built with a rigorous statistical methodology that can be validated by biological facts and whose analysis provides new scientific issues. It is the convergence between biologists' and statisticians' requirements. Finally, a good network is built by a good scientific and human collaboration network: so a good network is a network that makes everybody happy!

N. Villa-Vialaneix
IMT, Toulouse, France
L. Liaubet and M. SanCristobal
INRA, Toulouse, France
E-mail: magali.san-cristobal@toulouse.inra.fr

