Bayesian Integration of Genomic Data

Valeria Vitelli[1], **Øystein Sørensen**[1], **Elja Arjas**[1], **Magne Thoresen**[1] **and Arnoldo Frigessi**[1]

[1] Department of Biostatistics, Institute of Basic Medical Sciences, University of Oslo, Norway

Aim of the work is building a Bayesian model for integrating various sources of genomic data. The Oslo2 data include the gene expressions of 397 individuals for 41960 genomic features. Moreover, 148 protein level measurements (total and phosporilated proteins) are also available for 325 of the patients. These patients are women affected by breast cancer, and measurements come from tumor tissues. The action of the pathology can be detected both in the protein and expression activity levels, and in the way the variables are connected to one another. Final aim of the study is to associate specific genomic modifications to breast cancer subtypes.

The national cancer institute Pathway Interaction Database (PID) [3] contains human pathways composed of signaling and regulatory events: these pathways provide the physiological link between expression and protein levels, according to a specific signaling or regulatory task. Some of the pathways are known to be highly affected by breast cancer. However, little is known about modifications induced by the pathology on the network structure. The PID contains 137 pathways, including four different types of variables: small molecules, RNA, proteins, and complexes. We selected the *Class I PI3K signaling events mediated by Akt* pathway, due to its relevance in breast cancer.

A key step is to trasform this biological pathway into a probabilistic graphical model. This means creating a node for each genomic variable of each gene in each patient, which corresponds to a latent variable to be estimated. Then, all the nodes are connected according to the PID. When a protein reaches its active state, i.e. phosphorylates, a corresponding latent variable is created, and connected to the original protein and to all other variables involved in the phosphorylation process. Not all proteins have these active states, but some of them do. Some also have a number of different active states. Due to the fact that they represent active states, the phosphorylated variables are the most interesting to detect modifications induced by breast cancer. The first innovative aspect of our modeling strategy is that activity states are closely related to the biological structure of human pathways (this is the most relevant difference from the Paradigm strategy proposed by [4], the only other pathway-based genomic data integration approach proposed so far).

In our modeling framework, the physical modifications of the proteins (phosphorylation) are explicitly represented in the graphical model, and the same is true for translocations. We then model the effect of parent nodes on each

child node via a nonlinear monotone regression function, thus estimating from the data the relationships among genomic variables. The relationship is assumed to be monotone, due to the fact that typical interactions among molecules are either activation or inhibition. Due to the complexity of the network, we prefer a non-parametric Bayesian monotone regression approach [1]. Finally, we also include a Gaussian measurement error model linking each latent variable to the corresponding measurement.

There are two futher issues that still need to be mentioned: clustering, and pathway estimation. Concerning the former, since breast cancer patients are affected by different pathology subtypes, it has to be expected that the modifications induced by cancer act differently in different patients. Hence, we introduce in the model a clustering structure, and allow different connections (monotone regressions) to be estimated within groups. Moreover, we expect the biological pathway itself to be affected by cancer: the PID, in fact, contains physiological pathways only. We thus propose two strategies to handle this issue: either we estimate the pathway from the data, using a score-based Bayesian learning method [2], or we enrich the PID-based network with further consistent links, and then a posteriori evaluate their relevance.

Keywords: Integrative Genomics, Bayesian methods, Clustering, Pathways, Breast cancer.

References

[1] O. Saarela and E. Arjas (2011). A Method for Bayesian Monotonic Multiple Regression *Scand. Journ. Stat. 38*, 499–513.

[2] M. Scutari (2010). Learning Bayesian Networks with the bnlearn R Package *Journ. Stat. Soft.* **35** *3*, 1–22.

[3] C.F. Schaefer, K. Anthony, S. Krupa, J. Buchoff, M. Day, T. Hannay and K.H. Buetow (2009). PID: the Pathway Interaction Database *Nucleic Acids Research 37*, 674–679.

[4] C.J. Vaske, S.C. Benz, J.Z. Sanborn, D. Earl, C. Szeto, J. Zu, D. Haussler and J.M. Stuart (2010). Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM *Bioinf. 26*, 237–245.