Probabilistic dependency models for data integration in functional genomics

Leo Lahti (1) and Samuel Kaski (2)

(1) University of Helsinki, Department of Veterinary Bioscience, Finland (2) Aalto University, Department of Information and Computer Science, Adaptive Informatics Research Center and Helsinki Institute for Information Technology HIIT
leo.lahti@iki.fi, samuel.kaski@hiit.fi

Abstract. *Co-occurring genomic observations are increasingly available in biomedical studies, providing complementary views to genome function. Integrative analysis of these data sources can reveal dependencies and interactions that cannot be detected based on individual data sources. Prior information of the application domain can guide the search for novel multi-view biomarkers that have potential diagnostic, prognostic and clinical relevance. We propose an integrative analysis framework based on regularized probabilistic canonical correlation analysis with particular applications in cancer gene discovery and discuss other biomedically relevant extensions to the model.*

Keywords: canonical correlation analysis, data integration, dependency modeling, functional genomics

1 Introduction

Complementary genomic observations of gene- and micro-RNA expression, DNA copy number, and methylation status are increasingly available in biomedical studies public repositories such as the Cancer Genome Atlas [1]. Analysis of statistical dependencies between different functional layers of the genome allows the discovery of regularities and interactions that are not seen in individual data sets. For instance, integrative analysis of gene expression and copy number measurements can reveal cancer-associated chromosomal regions with potential clinical relevance. Variants of probabilistic canonical correlation analysis (CCA) [2] provide an intuitive framework for data integration in functional genomics that can deal with the uncertainties associated with small sample sizes common in biomedical studies and provide tools to guide dependency modeling through Bayesian priors [3]. We apply these models to detect and characterize functionally active chromosomal changes in gastric cancer.

2 Regularized dependency detection framework

Dependency between two data sources can be modeled by decomposing the observations into shared and data set specific components. Let us consider two sets

of co-occurring genomic observations, X, Y. The shared effects are described by a shared latent variable $\mathbf{z}$ whose manifestation in each data set is characterized by linear transformations W_x and W_y, respectively. Independent data set-specific effects are denoted by $\varepsilon_x, \varepsilon_y$. This gives

$$\begin{aligned} X &\sim W_x \mathbf{z} + \varepsilon_x \\ Y &\sim W_y \mathbf{z} + \varepsilon_y \end{aligned} \tag{1}$$

In standard probabilistic CCA [2], the shared latent variable $\mathbf{z}$ has standard multivariate normal distribution and the data-set specific effects are described by multivariate Gaussians with covariance matrices Ψ_x and Ψ_y, respectively. We incorporate domain-specific prior knowledge to focus on specific, biomedically relevant types of dependency. Biomedical screening studies often focus on particular types of regulation and unconstrained models easily lead to overfitting with small sample size. For instance, imposing particular structure on the marginal covariances could be used to data set specific prior information, and non-negativity constraints on W would emphasize positive regulatory relations. We show how constraining the relation between W_x and W_y help to model spatial dependencies in chromosomally local gene neighborghoods [3] and discuss other recent applications.

3 Detecting functionally active DNA mutations

DNA alterations are a key mechanism in cancer development. An important task in cancer studies is to distinguish so-called *driver* mutations from the less active *passengers*. Driver mutations that affect expression levels of the associated genes will contribute to dependencies between gene copy number and expression and detecting such regions will reveal potential candidate genes for cancer studies. Such dependencies are spatially constrained: probes with small chromosomal distance are expected to show similar changes in both data sources. This is encoded in the model by requiring that the transformation matrices W_x, W_y are similar. To enforce this we use a symmetric prior $W_x \sim N(W, \Sigma_w)$, $W_y \sim N(W, \Sigma_w)$. For simplicity, we use isotropic covariance matrix $\Sigma_w = \sigma I$, using σ to tune the similarity between W_x and W_y. With $\sigma \to \inf$ the transformations are uncoupled, yielding ordinary probabilistic CCA. Comparisons to another extreme, $\sigma \to 0$, which gives $W_x = W_y$ confirm that the regularized variant outperforms the unregularized model in cancer gene discovery.

To prioritize cancer-associated chromosomal regions, dependency is quantified within each gene neighborghood with a sliding window approach over the genome. The regions are sorted based on the dependency, quantified by the ratio of shared vs. data set-specific effects $\frac{Tr(WW^T)}{Tr(\Psi)}$, where $W = [W_x W_y]$ and Ψ is a block-diagonal matrix of the data set specific covariances Ψ_x, Ψ_y. A fixed dimensionality (window size around each gene) is used to obtain dependency scores that are directly comparable between the regions.

Figure 1**A** illustrates the dependency scores across chromosome arm 17q in gastric cancer patients [4], highlighting a known gastric cancer-associated chromosomal region. Genome-wide analysis of the dependencies confirms the overall cancer gene detection performance of the model [3] and shows favorable performance when compared to other recently proposed integrative approaches in cancer gene discovery (manuscript in preparation). The model parameters have straightforward interpretation: a maximum-likelihood estimate of the shared latent variable $\mathbf{z}$ indicates signal strength in each sample and W will describe probes that capture the shared signal (Fig. 1**B**) . The model detects rare copy number events that occur only in small subsets of patients and are manifested only in a subset of probes, which are important properties for cancer studies.

4 Conclusion

Modeling of dependencies can reveal regularities and interactions that are not seen in individual data sets. Regularized variants of probabilistic CCA provide efficient tools to investigate statistical dependencies between complementary genomic observations and to guide dependency detection through Bayesian priors. Freely available implementations of dependency detection models and application tools are available through CRAN[1] and BioConductor[2].

Acknowledgments. This work has been partially funded by TEKES (grant 40141/07).

5 References

References

1. The Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 455:1061–1068, 2008.
2. F. R. Bach and M. I. Jordan. A probabilistic interpretation of canonical correlation analysis. Technical report, Department of Statistics, University of California, Berkeley, 2005.
3. L. Lahti, S. Myllykangas, S. Knuutila, and S. Kaski. Dependency detection with similarity constraints. In *Proc. MLSP'09 IEEE International Workshop on Machine Learning for Signal Processing XIX*, pages 89–94, Piscataway, NJ, 2009.
4. S. Myllykangas, S. Junnila, A. Kokkola, R. Autio, I. Scheinin, T. Kiviluoto, M.-L. Karjalainen-Lindsberg, J. Hollmén, S. Knuutila, P. Puolakkainen, and O. Monni. Integrated gene copy number and expression microarray analysis of gastric cancer highlights potential target genes. *International Journal of Cancer*, 123:817–825, 2008.

[1] http://dmt.r-forge.r-project.org/
[2] http://www.bioconductor.org/packages/devel/bioc/html/pint.html

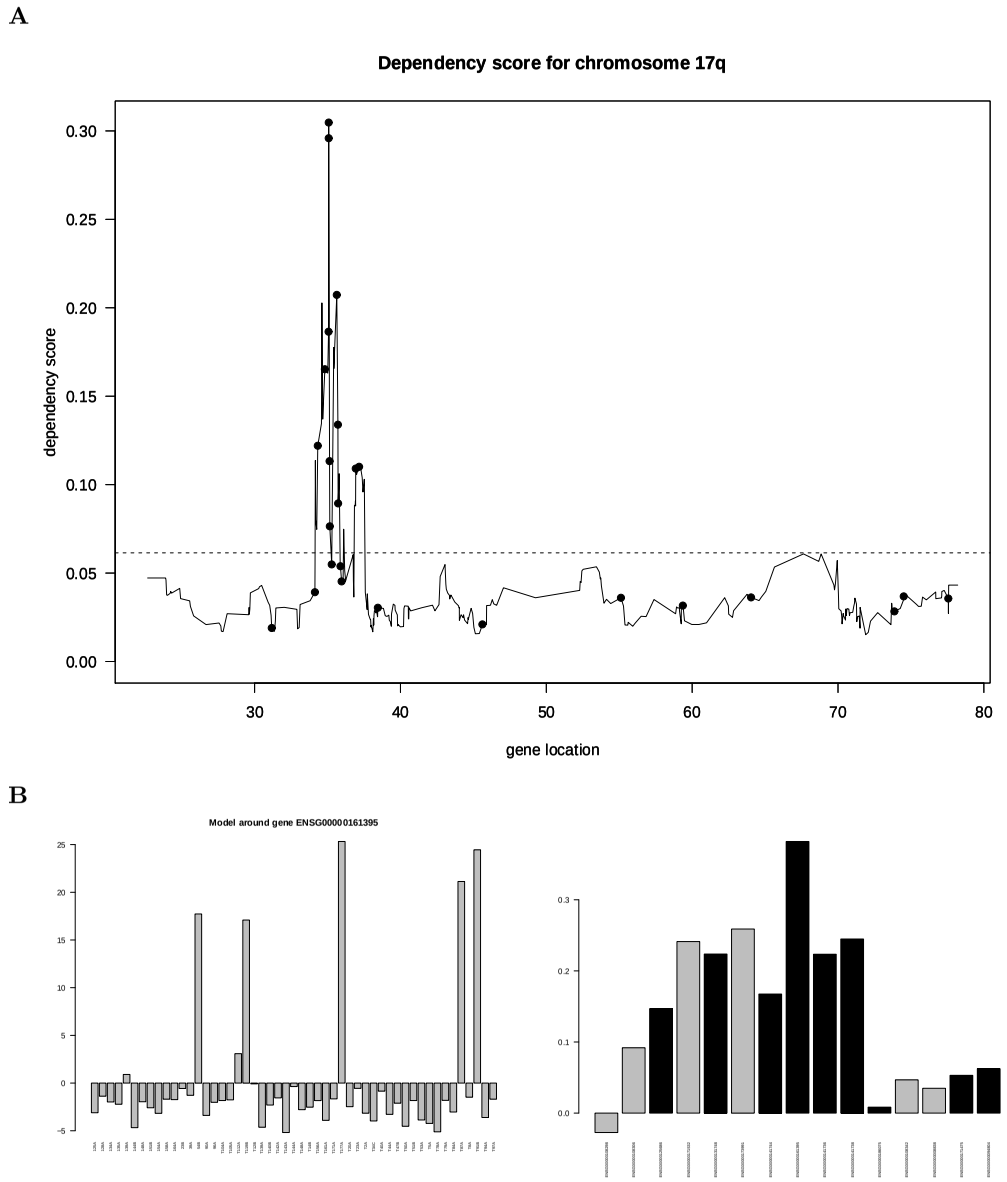

Fig. 1. A Dependency between gene copy number and expression across chromosomal arm 17q in gastric cancer. The black dots highlight known gastric cancer associated genes. **B** Samples and variable contribution to the dependencies around the gene with the highest dependency score between gene expression and copy number measurements in the region. The visualization highlights affected patients and genes.