Multiple hot-deck imputation for network inference from RNA sequencing data

Alyssa Imbert, Armand Valsesia, Claudia Armenise, Gregory Lefebvre, Pierre-Antoine Gourraud, Nathalie Viguerie, Nathalie Vialaneix

Description (150 words max.)

In recent years, high-throughput sequencing technology has become an essential tool for genomic studies. Furthermore, more and more studies include gene expression data measured by RNA-sequencing. Networks inferred from such data provide a global view of the relations existing between gene expression in a given transcriptomic experiment. However, the number of samples is still very small compared to the number of genes and that makes network inference an unrealistic objective for most RNA-seq experiments.

In order to take advantage of external information measured simultaneously to RNA-seq data, we propose a novel imputation method, hot-deck multiple imputation (**hd-MI**), that artificially increases the sample size and thus improves the reliability of network inference. **hd-MI** is able to impute samples which are missing for RNA-seq data but are observed on a secondary dataset (here, we used quantitative PCR data) by assuming gene expression similarity between samples that are similar for the secondary dataset.

Justification (250 words max.)

hd-MI was compared to direct inference and to other state-of-the art imputation methods on real world datasets. More stable and accurate networks were obtained, with an improvement in precision of edge prediction up to 30% of missing individuals.

Specifically focusing on a dataset acquired for the clinical-research project DiOGenes `http://www.diogenes-eu.org/`, we also found out that networks inferred by using **hd-MI** were coherent with previous findings based on other datasets and/or on a different subset of individuals. We were also able to discover novel links between genes that was coherent with functional knowledge.

More interestingly, since DiOGenes is a study based on a dietary intervention that aimed at investigating the effect of a low-calorie diet on adipose tissue in obese individuals, data have been acquired on the same individuals at two different time steps (before and after the diet). **hd-MI** was used to impute individuals missing in one time step before performing independent inference of two networks. The results showed an improved comparability between the networks inferred for the two steps of the nutritional intervention.

hd-MI is implemented in the R package **RNAseqNet**, released on CRAN and forthcoming research includes developing an inference method that can jointly estimate evolving network corresponding to distinct steps of a longitudinal analysis. Our purpose is to simultaneously handle missing individuals and explicitly model the relationships between expressions measured on the same individuals along the longitudinal study.