POSTER - SYSTEMS BIOLOGY AND MODELING

RECONSTRUCTION OF GENE REGULATORY NETWORKS WITH MULTIPLE DEPENDENCE AND CORRELATION METRICS IN GENE EXPRESSION DATA

Lucas Otávio Leme Silva (lucasotavio750@gmail.com)

Glaucia Maria Bressan (galbressan@gmail.com)

Alexandre Paschoal (alexandre.paschoal@rfi.ac.uk)

Fabricio Martins Lopes (fabricio@utfpr.edu.br)

Understanding fundamental biological processes, especially gene interactions, led to the concept of gene regulatory networks (GRNs), applied to cell differentiation, development, and disease progression. Analyzing these GRNs can shed light on the fundamental mechanisms of gene interactions, helping to unravel cellular functioning and the mechanisms underlying complex diseases. Assessing gene regulatory networks remains challenging due to the high dimensionality of gene expression data and limited samples. For example, the DREAM5 benchmark for Escherichia coli contains 4,511 genes and transcription factors, but only 805 samples, a scenario that makes it difficult to effectively apply deep learning-based models. Recent research has focused on developing sophisticated models to infer GRNs, often relying on correlation or dependence metrics for reconstruction. However, an in-depth analysis of different metrics is not addressed. Therefore, the objective of this work is to perform a comprehensive analysis of different dependency metrics and their respective behaviors, and also to propose a new metric that is the adaptation of mutual information using different entropies besides Shannon, such as Renyi and Tsallis. In total, 62 analyses were evaluated, covering linear, monotonic, nonlinear and nonmonotonic dependencies. The analyses were applied to pairs of genes and transcription factors present in the DREAM5 dataset, which contains 2,066 known interactions. The ability of each numerical metric to assign high scores to true interactions compared to false ones was evaluated, while statistical tests were only considered as significant or not (95% significance level). Preliminary bivariate results indicated that 486 of the 2,066 interactions were not identified by any numerical metric, and 572 were identified by only one. To find an ideal subset of numerical analyses, the number of correct relations identified had to be weighted and penalized by the number of metrics chosen. A planning for selecting subsets of metrics was formulated based on a scoring function, defined by: score(S) = alpha*Coverage(S) - beta(|S|), in which S is the subset of analyses selected. For preliminary results, alpha=1 and beta = relations/metrics = 2066/34 ~ 61 were used. The optimization of this function led to the selection of nine metrics that maximize the score. This subset was able to identify 1,152 of the 2,066 interactions. Despite finding about 55% of the relationships, the results still reveal important limitations in the detection of gene interactions based on dependency metrics. Among the main challenges are: the scarcity of samples compared to the high dimensionality of the data, the possibility of spurious correlations between unrelated variables, and the presence of local dependencies, observable only in specific subsets of experimental conditions. In contrast, all relationships were identified using statistical tests such as Hilbert Schmidt Independence, Heler Heler Gorfine, among others. However, the number of false positives is extremely high, but they can be good analyses to reduce the sample space. To advance the reconstruction of the genetic network, a new numerical metric that satisfies the following criteria is necessary: detection of arbitrary relationships (linear, non-linear, monotonic and non-monotonic), defined scale, robustness against false positives and ability to identify local dependencies.

Palavras-chave: gene regulatory networks; inference; correlation; metrics; entropy.