Poster J-22

Beyond classical statistics - A systems biology approach for pathway level analysis

Authors:
Sorin Draghici (*Karmanos Cancer Institute and Dept. of Comp. Science, Wayne State Univ.*)
Kashyap Amin (*Dept. of Comp. Science,*)
Arina Done (*Dept. of Comp. Science*)
Purvesh Khatri (*Dept. of Comp. Science*)

Short Abstract: A statistical approach is universally used to identify the most relevant regulatory pathways in high-throughput experiments. We show this approach is unsatisfactory, and can provide incorrect results. Using a systems biology approach, we developed a more powerful analysis that includes the classical statistics but also considers other crucial factors.

Long Abstract:
A common challenge in the analysis of genomics data is trying to understand the underlying phenomenon in the context of all complex interactions on various regulatory pathways. A statistical approach is universally used to identify the most relevant pathways in a given experiment. However, this approach only considers the number of differentially regulated genes on each pathway and completely ignores other important biological factors. We will show that the classical statistical analysis is unsatisfactory and often yields both false positives and false negatives. We show these limitations do not depend on the particular statistical model used but stem from the limited nature of the approach and its inability to consider very important biological factors. We will discuss how and why the limited numerical nature of this approach makes it unable to cope with the complex interactions and dependencies that characterize living organisms. To the best of our knowledge, this is the first time such intrinsic limitations will be shown for such a large category of analysis models.

Using a systems biology approach, we developed a more powerful impact analysis that extends the classical statistical approach by incorporating a number of crucial biological factors such as the magnitude of the expression change for each gene, the type and the position of the genes in the given pathways, the topology of pathway describing the regulatory interactions between genes, etc. To the best of our knowledge, none of the other approaches currently used for pathway analysis is able to integrate all these factors in a coherent model. Notably, the novel integrated model we propose is backwards compatible with the classical approach. When the limitations of the classical approach are forcefully imposed (e.g., ignoring the magnitude of the measured expression changes or ignoring the regulatory interactions between genes), the impact analysis reduces to the classical approach and yields exactly the same results. Both the backward coherency with the classical approach and the much expanded capabilities of this analysis make it likely to represent a breakthrough in this area.

Currently, all signalling pathways for human, mouse and rat are downloaded from KEGG and

stored in a relational database for efficient querying and retrieval. In order to calculate the impact factor for a given pathway, the pathway database is queried to retrieve all the genes and gene interactions in the pathway, and a graph data structure for this pathway is created. The genes are represented as nodes, and the gene interactions as edges of the graph. The user-provided normalized fold changes are mapped on the pathway graph and used to calculate some suitably defined gene perturbation factors. Once the perturbation factors of all genes in a given pathway are calculated, an impact factor is calculated for each pathway. The impact factor of each pathway is then used as a score to assess the impact of a given gene expression data set on all pathways (the higher the impact factor the more relevant the pathway in the given condition).

We use several real data sets to show that the classical analysis produces both false positives and false negatives while the impact analysis provides biologically meaningful results. We also discuss a few interesting particular cases of impact analysis. Some of these cases illustrate how this analysis reduces to the classical approach when certain restrictions are imposed; others clarify the role of some parameters and illustrate how the analysis can be refined to fit specific organisms or conditions.