Breaking the Power Law Curse: A Computational Bridge to Experiments (COMBREX)

Simon Kasif

Basic publication statistics show that the allocation of research attention to genes is far from uniform, following instead a power law distribution. Most genes have few or no citations and few genes are associated with a large number of references. This holds for human genes as well as genes in microbial organisms. While this publication bias could reflect an essential biological or medical significance associated with some genes it could also suggest that research tends to be clustered in silos and there could be benefits to breaking the "power law curse" by guiding experiments to produce the broadest predictive understanding of biology.

While many computational advances have been made in computational gene function prediction, few of these predictions are tested experimentally. We describe the initial steps we have taken to build a large consortium of computational researchers that will participate in populating the COMBREX database with predictions that will made available to experimental testing by participating biochemists. These efforts complement the parallel effort to build a synergistic experimental community. The COMBREX project directly funds experimental testing of predictions that are prioritized based on their overall utility.

COMBREX also attempts to produce a novel market model bridging between computational predictions and their experimental validation. A single experiment is obviously not sufficient to produce a computational model of gene function for a large protein family and its validation. We discuss the COMBREX project as a first of a kind community implementation of Active Learning paradigm that aims to drive predictive sciences by cascades of experiments testing hypotheses that are expected to have the highest impact on the overall accuracy of predictive models used in the field. This approach is expected to change the current culture where experimental data generation is followed by analysis and model building resulting in models that are better at explaining the data instead of maximizing their predictive power.