ATELIER « MÉTHODES STATISTIQUES COMPUTATIONNELLES EN GÉNOMIQUE ET EN BIOLOGIE SYSTÉMIQUE »

18–22 AVRIL 2011

WORKSHOP ON "COMPUTATIONAL STATISTICAL METHODS FOR GENOMICS AND SYSTEMS BIOLOGY"

APRIL 18–22, 2011

Querying genomic databases : refining the connectivity map

Mark R. Segal *
mark@biostat.ucsf.edu

The advent of high throughput biotechnologies, that can efficiently measure gene expression on a global basis, has led to the creation and population of correspondingly rich databases and compendia. Such repositories have the potential to add enormous scientific value beyond that provided by individual studies which, due largely to cost considerations, are typified by small sample sizes. Accordingly, substantial effort has been invested in devising analysis schemes for utilizing gene expression repositories. Here, we focus on one such scheme, the connectivity map, that was developed with the express purpose of identifying drugs with putative efficacy against a given disease, where the disease in question is characterized by a gene expression signature. In view of the enormous costs and poor success rates of established drug development pipelines, the promise seemingly demonstrated by early use of the connectivity map is of profound importance.

The success of the connectivity map is belied by its simplicity. The aforementioned signature serves as a query which is applied to a customized database of (differential) gene expression experiments designed to elicit response to a wide range of drugs, across of spectrum of concentrations, durations, and cell lines. Such application is effected by computing a per experiment score that measures "closeness" between the signature and the experiment. Top scoring experiments, and the attendant $\mathrm{drug}(s)$, are then deemed relevant to the disease underlying the query. Inference supporting such elicitations is pursued via resampling. In this talk, we revisit two key aspects of the connectivity map implementation. Firstly, we develop new approaches to measuring closeness for the common scenario wherein the query constitutes an ordered list. These involve using metrics proposed for analyzing partially ranked data. Secondly, we advance an alternate inferential approach based on generating empiric null distributions that exploit the scope, and capture dependencies, embodied by the database. Using these refinements we illustrate select results from a comprehensive re-evaluation of connectivity map findings.

*Department of Epidemiology & Biostatistics University of California, San Francisco, 185 Berry Street, Lobby 5, Suite 5700, San Francisco, CA 94107-1762, USA.