Algorithmic Challenges for Systems-Level Correlational Analysis: A Tale of Two Datasets*

Michael A. Langston

Department of Computer Science, University of Tennessee, Knoxville, TN 37996–3450, USA

I will discuss novel algorithmic, combinatorial and correlational tools for the analysis of complex natural systems. A pair of illustrative but widely divergent applications will be described. Despite huge differences in data acquisition methodologies, the algorithmic missions for both problems are similar, and help to highlight the rich interplay between data quality and effective computation.

The first application centers on determining the effects of environment on man. As a case study, we search for biological pathways relevant to the human allergic response. We exploit well-designed studies and quantitative data generated with state-of-the-art technologies. We extract putative relationships from the simultaneous expression of vast numbers of genes, under the premise that genes encoding proteins functioning in a common pathway often exhibit correlated levels of expression. Thus the identities and ontologies of these genes can be used to pinpoint existing and assimilate new functional pathway elements. Armed with advanced technologies and high-quality data, we seek to elucidate genetic components relevant to allergic rhinitis, asthma and eczema.

The second application focuses on the rather complementary problem of determining the effect of man on environment. As a case study, we analyze quantifiable variables of significance to oceanic ecosystems. These variables encompass a huge variety of biotic and abiotic factors, and tend to possess differing periodicities and other diverse properties. Only heuristic experimental designs and incomplete and sometimes dubious historical data is available. We labor to uncover temporal, spatial and other meaningful patterns on an immense scale, and to shed light on inflection points, putative regime changes and other complex relationships. Data quality and missing or corrupted values are significant, as is the mining of information at multiple levels of granularity. Armed with powerful technologies but highly challenging data, we seek to establish dependencies upon which we can draw conclusions about the impact of man and other agents upon the sea.

* This research has been supported in part by the U.S. National Institutes of Health under grants 1-P01-DA-015027-01, 5-U01-AA-013512 and 1-R01-MH-074460-01, by the U.S. Department of Energy under the EPSCoR Laboratory Partnership Program, by the Australian Research Council, and by the European Commission under the Sixth Framework Programme.

F. Dehne, J.-R. Sack, and N. Zeh (Eds.): WADS 2007, LNCS 4619, p. 226, 2007.