

Abstract for an Invited Paper
for the MAR05 Meeting of
The American Physical Society

Genomic Signal Processing: Predicting Basic Molecular Biological Principles

ORLY ALTER, Department of Biomedical Engineering & Institute for Cellular and Molecular Biology, UT Austin

Advances in high-throughput technologies enable acquisition of different types of molecular biological data, monitoring the flow of biological information as DNA is transcribed to RNA, and RNA is translated to proteins, on a genomic scale. Future discovery in biology and medicine will come from the mathematical modeling of these data, which hold the key to fundamental understanding of life on the molecular level, as well as answers to questions regarding diagnosis, treatment and drug development. Recently we described data-driven models for genome-scale molecular biological data, which use singular value decomposition (SVD) and the comparative generalized SVD (GSVD). Now we describe an integrative data-driven model, which uses pseudoinverse projection (1). We also demonstrate the predictive power of these matrix algebra models (2).

The integrative pseudoinverse projection model formulates any number of genome-scale molecular biological data sets in terms of one chosen set of data samples, or of profiles extracted mathematically from data samples, designated the “basis” set. The mathematical variables of this integrative model, the pseudoinverse correlation patterns that are uncovered in the data, represent independent processes and corresponding cellular states (such as observed genome-wide effects of known regulators or transcription factors, the biological components of the cellular machinery that generate the genomic signals, and measured samples in which these regulators or transcription factors are over- or underactive). Reconstruction of the data in the basis simulates experimental observation of only the cellular states manifest in the data that correspond to those of the basis. Classification of the data samples according to their reconstruction in the basis, rather than their overall measured profiles, maps the cellular states of the data onto those of the basis, and gives a global picture of the correlations and possibly also causal coordination of these two sets of states.

Mapping genome-scale protein binding data using pseudoinverse projection onto patterns of RNA expression data that had been extracted by SVD and GSVD, a novel correlation between DNA replication initiation and RNA transcription during the cell cycle in yeast, that might be due to a previously unknown mechanism of regulation, is predicted.

(1) Alter & Golub, *Proc. Natl. Acad. Sci. USA* **101**, 16577 (2004).

(2) Alter, Golub, Brown & Botstein, *Miami Nat. Biotechnol. Winter Symp. 2004* (www.med.miami.edu/mnbws/alter-.pdf)