BICLUSTERING METHODS FOR RE-ORDERING DATA MATRICES IN SYSTEMS BIOLOGY, DRUG DISCOVERY AND TOXICOLOGY

Christodoulos A. Floudas
Princeton University, Department of Chemical Engineering
Princeton, NJ 08544
Phone: 001-609-258-4595; E-mail: floudas@princeton.edu

Key words: *biclustering, row/column-reordering, systems biology, TSP, MILP*

EXTENDED ABSTRACT

Biclustering has emerged as an important problem in the analysis of gene expression data since genes may only jointly respond over a subset of conditions. Many of the methods for biclustering, and clustering algorithms in general, utilize simplified models or heuristic strategies for identifying the ``best" grouping of elements according to some metric and cluster definition and thus result in suboptimal clusters.

In the first part of the presentation, we present a rigorous approach to biclustering, OREO, which is based on the Optimal RE-Ordering of the rows and columns of a data matrix so as to globally minimize the dissimilarity metric [1,2]. The physical permutations of the rows and columns of the data matrix can be modeled as either a network flow problem or a traveling salesman problem. The performance of OREO is tested on several important data matrices arising in systems biology to validate the ability of the proposed method and compare it to existing biclustering and clustering methods.

In the second part of the talk, we will focus on novel methods for clustering of data matrices that are very sparse [3]. These types of data matrices arise in drug discovery where the x- and y-axis of a data matrix can correspond to different functional groups for two distinct substituent sites on a molecular scaffold. Each possible x and y pair corresponds to a single molecule which can be synthesized and tested for a certain property, such as percent inhibition of a protein function. For even moderate size matrices, synthesizing and testing a small fraction of the molecules is labor intensive and not economically feasible. Thus, it is of paramount importance to have a reliable method for guiding the synthesis process to select molecules that have a high probability of success. In the second part of the presentation, we introduce a new strategy to enable efficient substituent reordering and descriptor-free property estimation. Our approach casts substituent reordering as a special high-dimensional rearrangement clustering problem, eliminating the need for functional approximation and enhancing computational efficiency [4, 5]. Deterministic optimization approaches based on mixed-integer linear programming can provide guaranteed convergence to the optimal substituent ordering. The proposed approach is demonstrated on a sparse data matrix (about 29% dense) of inhibition values for 14,043 unknown compounds provided by Pfizer Inc. It is shown that an iterative synthesis strategy is able to uncover a significant percentage of the lead molecules while using only a fraction of total compound library, even when starting from a mere 3\% of the total library space. In the third part of the presentation, we combine the strengths of integer linear optimization and machine learning to predict in vivo toxicities for a library of pesticide chemicals using only in vitro data. Our approach utilizes a biclustering method based on iterative optimal re-ordering [1,2] to identify biclusters corresponding to subsets of chemicals that have similar responses over distinct subsets of the in vitro assays. This enables us to determine subsets of experimental assays that are most likely to be correlated with toxicity, according to the in vivo data set. An optimal method based on integer linear optimization (ILP) for re-ordering *sparse* data matrices [3] is also applied to the in vivo dataset (21.3% sparse) in order to cluster endpoints that have similar lowest effect level (LEL) values, where it is observed that endpoints are grouped according to similar physiological attributes. Based upon the clustering results of the in vitro and in vivo data sets, logistic regression is then utilized to (a) learn the correlation between the subsets of in vitro data and the in vivo responses, and (b) subsequently predict the toxicity signatures of the chemicals. Our approach aims to find the highest prediction accuracy using the minimum number of in vitro descriptors.