Optimal Methods for Re-Ordering Data Matrices in Systems Biology and Drug Discovery Applications

P. A. DiMaggio[1], S. R. McAllister[1], and C. A. Floudas[1]

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. 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.

Another challenging problem in clustering is the rearrangement of data matrices that are very sparse. 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. 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.

[1] Department of Chemical Engineering
Princeton University
Princeton, NJ, 08544, USA
floudas@titan.princeton.edu