OPEN ACCESS Freely available online PLoS COMPUTATIONAL BIOLOGY

MCAM: Multiple Clustering Analysis Methodology for Deriving Hypotheses and Insights from High-Throughput Proteomic Datasets

Kristen M. Naegle[1,2], Roy E. Welsch[3], Michael B. Yaffe[1,2,4], Forest M. White[1,2], Douglas A. Lauffenburger[1]*

1 Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts, United States of America, **2** Koch Institute for Integrative Cancer Research, Massachusetts Institute of Technology, Cambridge, Massachusetts, United States of America, **3** Sloan School of Management, Massachusetts Institute of Technology, Cambridge, Massachusetts, United States of America, **4** Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts, United States of America

Abstract

Advances in proteomic technologies continue to substantially accelerate capability for generating experimental data on protein levels, states, and activities in biological samples. For example, studies on receptor tyrosine kinase signaling networks can now capture the phosphorylation state of hundreds to thousands of proteins across multiple conditions. However, little is known about the function of many of these protein modifications, or the enzymes responsible for modifying them. To address this challenge, we have developed an approach that enhances the power of clustering techniques to infer functional and regulatory meaning of protein states in cell signaling networks. We have created a new computational framework for applying clustering to biological data in order to overcome the typical dependence on specific *a priori* assumptions and expert knowledge concerning the technical aspects of clustering. Multiple clustering analysis methodology ('MCAM') employs an array of diverse data transformations, distance metrics, set sizes, and clustering algorithms, in a combinatorial fashion, to create a suite of clustering sets. These sets are then evaluated based on their ability to produce biological insights through statistical enrichment of metadata relating to knowledge concerning protein functions, kinase substrates, and sequence motifs. We applied MCAM to a set of dynamic phosphorylation measurements of the ERRB network to explore the relationships between algorithmic parameters and the biological meaning that could be inferred and report on interesting biological predictions. Further, we applied MCAM to multiple phosphoproteomic datasets for the ERBB network, which allowed us to compare independent and incomplete overlapping measurements of phosphorylation sites in the network. We report specific and global differences of the ERBB network stimulated with different ligands and with changes in HER2 expression. Overall, we offer MCAM as a broadly-applicable approach for analysis of proteomic data which may help increase the current understanding of molecular networks in a variety of biological problems.