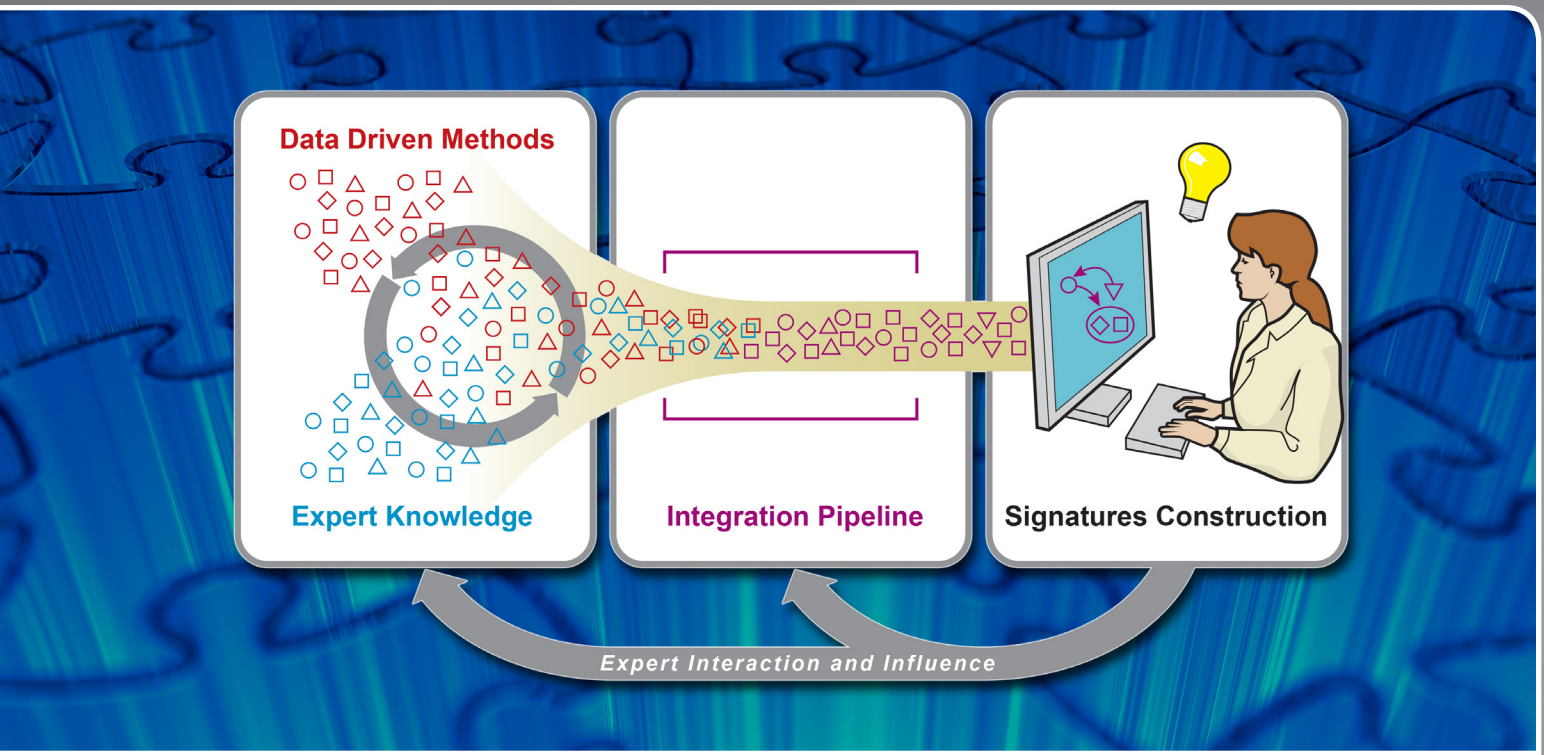
Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965



Experts Inundated with Data: The Biomarkers Problem

Our objective is to develop a powerful tool for successfully applying expert knowledge to data-driven analysis of large complex datasets to extract the most robust signatures.

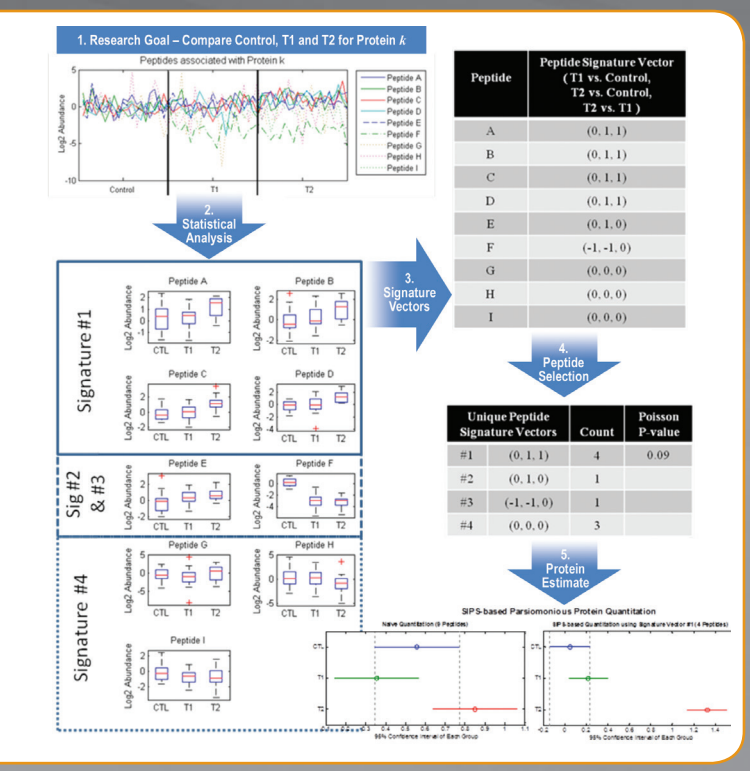
CHALLENGE

A primary challenge facing biomarker scientists is the need to integrate expert-driven and data-driven approaches to biomarker discovery. Relying exclusively on either approach has proven unsatisfactory and many recent meta-studies on biomarkers and signatures derived from statistical analysis of high-throughput data have shown that these do not generalize to larger, more complete datasets.

APPROACH/METHODS

Expert-driven and statistical, data-driven approaches are being combined in a domain-independent workflow for biosignature identification that will provide expert input into preprocessing and optimize clustering algorithms to integrate data-based relationships between variables with ontological relationships for classification. Our pipeline combines similarity matrices based on distances between variables, abundance levels, and ontological distance. The combined matrix is used to divide the variables into related clusters, which are used as input to a Bayesian data classification algorithm. Optimization is performed on this process to derive the set of clusters that provides the optimal discrimination between classes, such as disease states. The resulting model (or signature) has better classification performance and is more robust across different datasets.

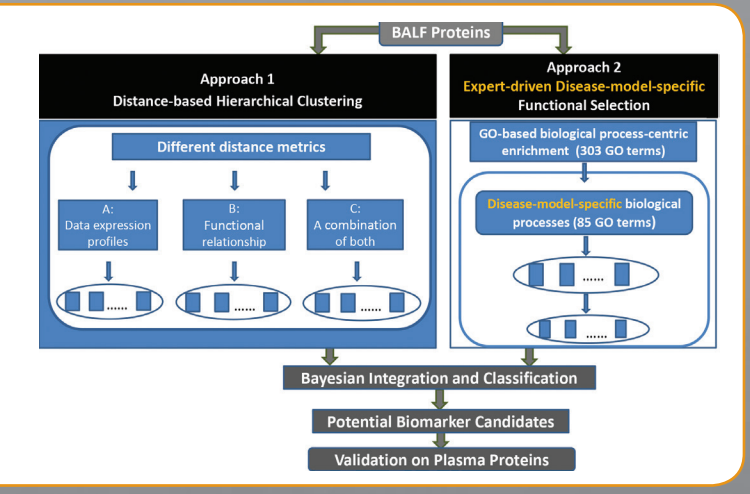
Genes and proteins are organized in biological systems into modules that operate coordinately to carry out particular functions. Biological experts in the community have annotated genes and proteins with these functions providing a community-based compilation of expert knowledge in the form of a biological ontology. A significant problem in derivation of biosignatures of disease is that patterns determined using solely statistical methods fail to generalize to other sets of patients. Our solution to this problem is to



Peptide	Peptide Signature Vector (T1 vs. Control, T2 vs. Control, T2 vs. T1)
A	(0, 1, 1)
B	(0, 1, 1)
C	(0, 1, 1)
D	(0, 1, 1)
E	(0, 1, 0)
F	(-1, -1, 0)
G	(0, 0, 0)
H	(0, 0, 0)
I	(0, 0, 0)

Unique Peptide Signature Vectors		Count	Poisson P-value
#1	(0, 1, 1)	4	0.09
#2	(0, 1, 0)	1	
#3	(-1, -1, 0)	1	
#4	(0, 0, 0)	3	

Overview of the Statistically Informed Peptide Selection Approach for Protein Quantitation Developed Under this Project



Application of Our Pipeline to Identify Biosignatures of Chronic Obstructive Pulmonary Disease

use similarity between genes or proteins based on patterns of expression and on functional similarity using biological ontology. These two sets of relationships are merged and then used to divide the variables into clusters related by both expression pattern and function. Finally, a Bayesian classification algorithm is used to train a classifier based on these clusters. This approach is powerful because it identifies modules that serve as biosignatures; the modules are likely to be more robust across different datasets because they represent functional groups. To identify biosignatures, the pipeline is optimized for multiple parameters and the set of clusters with the best classification accuracy is selected. Further feature selection approaches can be used to refine the biosignature.

IMPACT/BENEFIT

The abstract principles developed can be applied to any type of complex, high-density data including disease detection and prognosis, problems in environmental remediation, national defense, and energy production. The signatures produced by combining expert insight with data will be more robust and more interpretable than those relying on one source alone. The pipeline is easily adaptable to other domains and can incorporate other sources of relationships between variables.

ACCOMPLISHMENTS/OUTCOMES

We developed a pipeline to merge data- and expert-driven relationships between variables to improve classification. We identified proteomic signatures of chronic obstructive pulmonary disease. Additionally, we cross-validated a pipeline on large cancer datasets to identify signatures of chemo resistance in ovarian cancer. To this end, we have published or submitted four papers.

Jason McDermott
Pacific Northwest National Laboratory
jason.mcdermott@pnnl.gov | (509) 372-4360

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle *Since 1965*

U.S. DEPARTMENT OF ENERGY