PNNL-29450

Machine Learning for Rapid Biomarker Discovery via Image-Omic Fusion

November 2019

Bryan A Stanfill
Lisa M Bramer
Tao Liu
Sarah M Akers

U.S. DEPARTMENT OF ENERGY

Prepared for the U.S. Department of Energy
under Contract DE-AC05-76RL01830

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor Battelle Memorial Institute, nor any of their employees, **makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights**. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or Battelle Memorial Institute. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

PACIFIC NORTHWEST NATIONAL LABORATORY
operated by
BATTELLE
for the
UNITED STATES DEPARTMENT OF ENERGY
under Contract DE-AC05-76RL01830

Printed in the United States of America

**Available to DOE and DOE contractors from
the Office of Scientific and Technical
Information,
P.O. Box 62, Oak Ridge, TN 37831-0062**
www.osti.gov
**ph: (865) 576-8401
fox: (865) 576-5728
email: reports@osti.gov**

**Available to the public from the National Technical Information Service
5301 Shawnee Rd., Alexandria, VA 22312
ph: (800) 553-NTIS (6847)
or (703) 605-6000
email: info@ntis.gov
Online ordering: http://www.ntis.gov**

PNNL-29450

Machine Learning for Rapid Biomarker Discovery via Image-Omic Fusion

November 2019

Bryan A Stanfill
Lisa M Bramer
Tao Liu
Sarah M Akers

Prepared for
the U.S. Department of Energy
under Contract DE-AC05-76RL01830

Pacific Northwest National Laboratory
Richland, Washington 99354

Abstract

In this project we proposed to accelerate biomarker discovery and improve disease prediction by combining image and omic data using a single, unified modelling approach. We hypothesized that much richer information could be derived from both the image and omic data when they are analyzed together. Using imaging mass spectrometry (IMS) data collected from mice we fused image and omic data to increase the resolution of the IMS data. Using human data provided to us by Oregon Health and Science University, we developed an initial modeling framework that can increase the accuracy of both imaging and MS datasets.

Acknowledgments

This research was supported by the Earth and Biological Sciences Directorate (EBSD) Mission Seed, under the Laboratory Directed Research and Development (LDRD) Program at Pacific Northwest National Laboratory (PNNL). PNNL is a multi-program national laboratory operated for the U.S. Department of Energy (DOE) by Battelle Memorial Institute under Contract No. DE-AC05-76RL01830.

1.0 Research

The goal of this project was to develop a machine learning method to fuse image and omic information to improve biomarker discovery and improve disease prediction accuracy. We did this with two different data sets. The first was furnished by Kristin Burnum-Johnson (manuscript in progress). The data set consists of images and proteomic measurements collected from mouse samples. Their analysis focused on identifying the peptides that were differentially expressed in the different tissue types present in the sampled tissue. The proteomic measurements were taken from sample squares that measured 50 microns on each side. In practice, they would like to be able to resolve the proteomic maps at a higher resolution.

To increase the resolution of the peptide data, we used image processing techniques to turn the one-dimensional grayscale image of the mouse sample into separate channels that represent different features of the image, e.g., boundaries, tissue categories, and gradient changes, in order to derive more information from the image. When then appended the proteomic information in order to correlate image features with the map of protein abundances. Because the image data are available at the pixel level, we were able to increase the resolution of the peptide information to the pixel level as well. There are >193K pixels in each of the 50-micron squares from which the peptide information was derived.

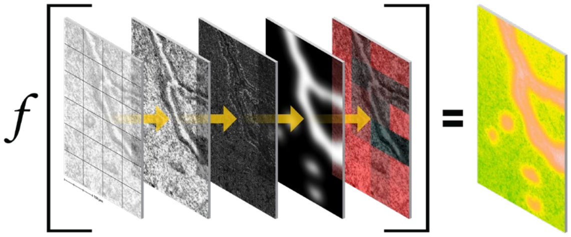
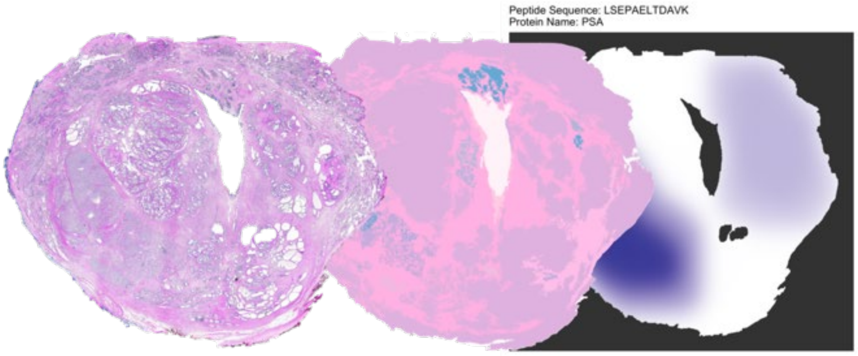


Illustration of how the image-omic methods developed in this project were used to increase the resolution of IMS data from 25 μm cells to individual pixels.

An illustration of the proposed method is above in which the channels in the image are represented by the first four images inside of the brackets while the peptide information is the final image with the red and black squares. The "f" represents our method of fusing all five images and the final image on the right-hand side of the equal sign contains the pixel level predictions for the peptide used to train the model. Again, the peptide abundances are plotted at the pixel level.

After developing this initial method on the mouse data, we also analyzed human data provided to us by our collaborators at Oregon Health and Science University (OHSU). They sent us three tissue samples along with images of the tissue with health and tumor areas identified. The normal and tumor tissues (along with some surrounding area) were analyzed using PRISM-SRM which generated peptide abundances for 64 peptides in each tissue type.



The image and omic data provided to use from OHSU as a part of this project were analyzed using image processing (middle) and statistical methods (right).

The peptide data were initially analyzed on their own in order to identify if any of them are differentially expressed in the tumor versus normal tissue types. Twenty-five of the peptides showed significantly different abundances in the tissue types. The image data were then analyzed in order to identify image features that correlate with normal and tumorous tissues. The most important image feature in determining if a region of the image represents a healthy or tumorous area is the darkness and density of the region. Because we only have three tissue samples and four tissue types, we are not currently able to fuse the image and omic information in a meaningful way. However, these initial results will be used in an NIH R01 grant proposal in order to secure funding to generate more data for the fusion component of the method.

Pacific Northwest National Laboratory

902 Battelle Boulevard
P.O. Box 999
Richland, WA 99354
1-888-375-PNNL (7665)

www.pnnl.gov