Rapid Bioinformatics Prototyping Based on a Public Application Programming Interface

Ourcon 2019, #79

J.H. Kobarg[1], N. Verbeeck[2], D. Lachmund[3], J. von Schroeder[3], S.O. Deininger[1], T. Moerman[2], S. Schiffler[1], J. Singe[1], D. Trede[1], M. Claesen[2], T. Boskamp[1,3]
[1]SCiLS / Bruker Daltonik GmbH, Bremen, Germany; [2]Aspect Analytics NV, Genk, Belgium; [3]University of Bremen, Center for Industrial Mathematics, Bremen, Germany

Introduction

Modern mass spectrometry imaging (MSI) instrumentation enables high spatial resolution and mass resolving power, resulting in large binary data sets.

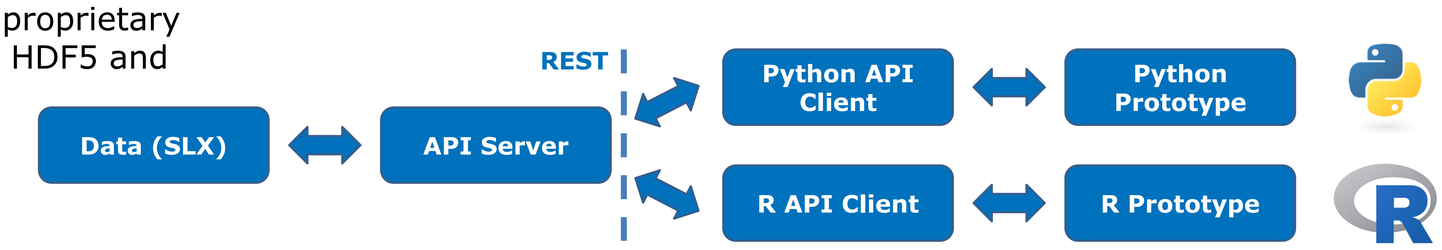
Existing open data formats, including imzML (Römpp et.al., 2011) are not always suitable for fast data access and high performance computing.

We present an approach for rapid prototyping applied to large MSI data through an open application programming interface (API).

Data is stored in the proprietary SLX format based on HDF5 and SQLite.

The API is implemented as a Representational State Transfer (REST) server, accessed through API clients in Python and R.

API server, as well as Python and R clients are currently in beta. Client implementations for Matlab and C++ are in preparation.



Prototype 1:
Unsupervised non-linear dimensionality reduction

Four different methods for unsupervised dimensionality reduction were implemented in Python and evaluated on three test datasets:

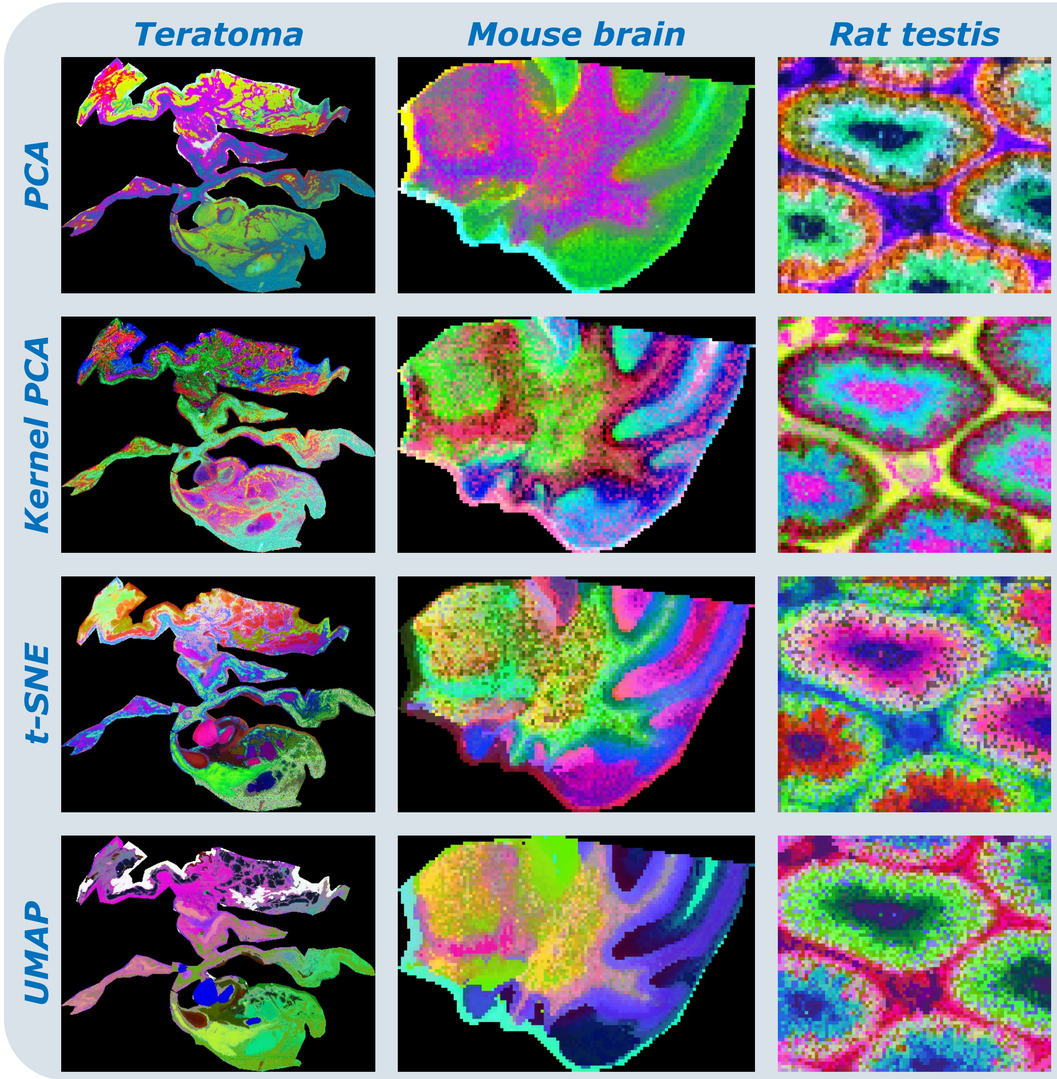
a. Teratoma: rapifleX, m/z 600–3200, 50 µm, 89000 spots, 53300 bins (*courtesy of J. Kriegsmann, R. Casadonte, Proteopath, Trier*)

b. Mouse brain: rapifleX, m/z 600–3200, 50 µm, 6000 spots, 33400 bins

c. Rat testis: solariX, m/z 150–3000, 50 µm, 5700 spots, 2.9 mio bins

Dataset	PCA	Kernel PCA*	t-SNE	UMAP
Teratoma	4.47s	26.6s	10h16’	8’36s
Mouse brain	0.32s	16.6s	6’26s	30s
Rat testis	0.30s	18.9s	6’20s	21s

Execution times on standard desktop hardware, no GPU acceleration.
**Kernel PCA based on a naïve, non-optimized implementation. With optimization, performance close to PCA expected.*



All methods were performed on 5000 m/z bins with highest intensity in mean spectrum.

Non-linear histogram equalization was used on each color channel separately.

PCA: First three principal components calculated and visualized

Kernel PCA: Approximate kernel PCA with cosine kernel

t-SNE: Barnes-Hut approximation, Euclidean distance, perplexity 30

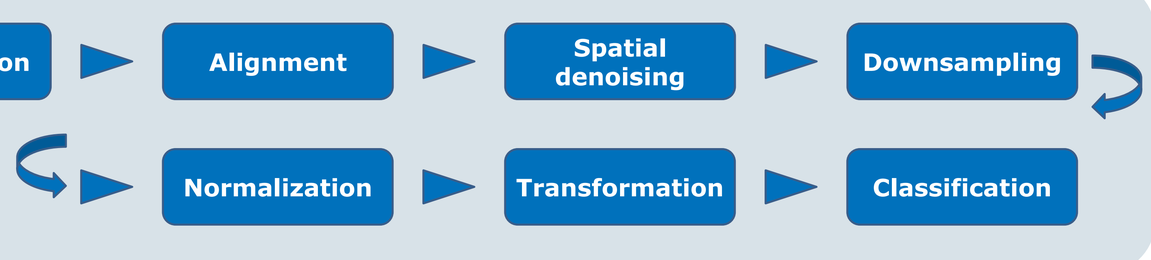
UMAP: Uniform Manifold Approximation and Projection, *Smets et.al, Anal. Chem. 2018*

Embedding: Low dimensional embedding for rat testis data.

Prototype 2:
Supervised machine learning pipeline for FFPE tissue typing

An optimized preprocessing pipeline for MALDI MSI based tumor typing was developed and evaluated in R.

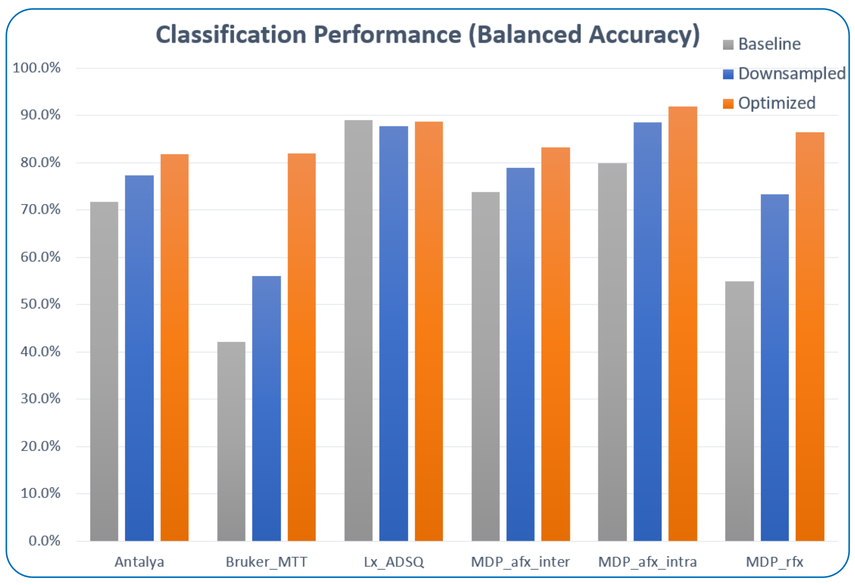
Machine learning pipeline from *Lachmund et.al., ASMS 2019*, including non-linear intensity profile normalization, mass alignment and spectral downsampling (*Boskamp et.al., ASMS 2018*), as well as spatial denoising and intensity log-transform.



Evaluation was performed on a panel of six clinical tumor typing and subtyping benchmark tasks, acquired from 25 TMAs, 2031 cores and 1410 patients total

Task	Instrument	Description
Antalya	autoflex	• Four tumor entities, 8 TMAs • Lung, pancreas, colon, breast
Bruker MTT	rapiflex	• Six tumor entities on one TMA • Five measurements in four labs • Training and test data from different SOP’s
Lx ADSQ	autoflex	• Eight TMAs with mix of adeno- and squamous cell carcinoma
MDP afx inter	autoflex	• Breast, ovary tumors, 5 TMAs • Measured in two labs • Inter-lab cross-validation
MDP afx intra	autoflex	• Same as above, but intra-lab cross-validation
MDP rfx	rapiflex	• Breast, ovary tumors, 5 TMAs • Single lab

Optimized pipeline significantly increases classification performance across five of six tasks. Lower gains obtained with mass alignment and downsampling alone (blue bars).



Summary

- API with Python and R client implementations enables rapid prototyping and evaluation of complex data analysis algorithms
- API is demonstrated by two prototypes implementing supervised and unsupervised machine learning methods

Bioinformatics