Unbiased Disease Stratification within the IMI U-BIOPRED Severe Asthma Programme Using Topological Data Mining

Dominic Burg [1,2], Doroteya Staykova[1], Pek Lum[3], Xian Yang[4], Yike Guo[4], Anthony Rowe[5], Ratko Djukanović[2], Paul Skipp[1] and the U-BIOPRED consortium

1. Centre for Proteomic Research, University of Southampton, UK. **2.** NIHR Southampton Respiratory Biomedical Research Unit, Southampton University Hospital, UK. **3.** Ayasdi Inc, USA. **4**. Imperial College, London, UK. **5.** Janssen R&D, UK

Background and approach

The U-BIOPRED consortium is an EU-wide collective of academics (20 institutions), biopharma industry (12), SME's (3), and patient organisations (6) working collaboratively to improve understanding of severe asthma. Representing the largest study cohort recruited for this disease, a variety of sample types are currently being analysed in parallel using a range of 'omics technologies to map molecular and clinical phenotypes of severe asthma in an unbiased manner. The heterogeneity of the disease, combined with the complexity of the study cohort (e.g. participants on a combination of medications, and varying co-morbidities), the range of biofluids and tissues analysed (each with corresponding challenges), and the multiple analytes being measured (e.g. lipids, proteins, mRNA) have necessitated the development of multiple data analysis pipelines to mine these complex datasets. One of the approaches used by the consortium is Topological Data Analysis (TDA), implemented via the Ayasdi software platform. TDA generates topological networks (Figures 1&2) that allow the scientist to explore, condense, visualise and extract useful information from these complex and multi-modal data. The subsequent sections show examples of how the Ayasdi platform is being used to combine and analyse proteomic data produced in the U-BIOPRED study, particularly highlighting the utility of the approach in exploring the biology of the data and as an unbiased feature selection tool. The datasets used to construct the example analyses represent only a small fraction of the final UBIOPRED cohort. As such, all results and interpretations must be viewed as exploratory and illustrative of approach and may not be representative of any final outcomes of the study.

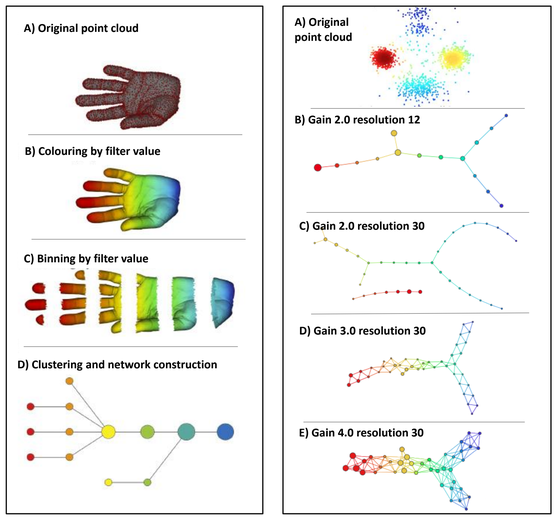


Figure 1(Left) Summary of the Ayasdi approach[1] **A)** A 3D object (hand) represented as a point cloud. This point cloud can also be generated through e.g. correlation analysis of a complex dataset **B)** A filter value is applied to the point cloud and the object is now coloured by the values of the filter function. **C)** The data set is binned into overlapping groups. **D)** Each bin is clustered and a network is built. **Figure 2** (right) Resolution and gain of graphs from : **A)** analysis of a point cloud . **(B&C)** Resolution alters the number of data points in each node **(D&E)** Gain alters the bin overlap and the subsequent number of edges between nodes

Exploring features of complex data from multiple sources

Proteomic data, obtained from the analysis of sputum supernatant and top-12 depleted serum using the data independent acquisition approach MS^E, were searched against the UNIPROT database for identity assignment and quantification using the Hi-3 method[2]. Each dataset was normalised and batch effect correction performed where necessary using modified ComBat[3] scripts. Proteomic data were subsequently aligned and clinical information were added as metadata, resulting in a complete dataset for 80 asthmatic participants. Missing values in all data types were not imputed and left as 'null'. Data were loaded into the Ayasdi platform, including proteins with up to 60% null values, and analysed with a normalised correlation metric and multi-dimensional scaling (MDS) lenses. The resultant graph was explored using the proteomic and clinical data, and groups selected using the clinical meta-data (Figures 3 & 4).

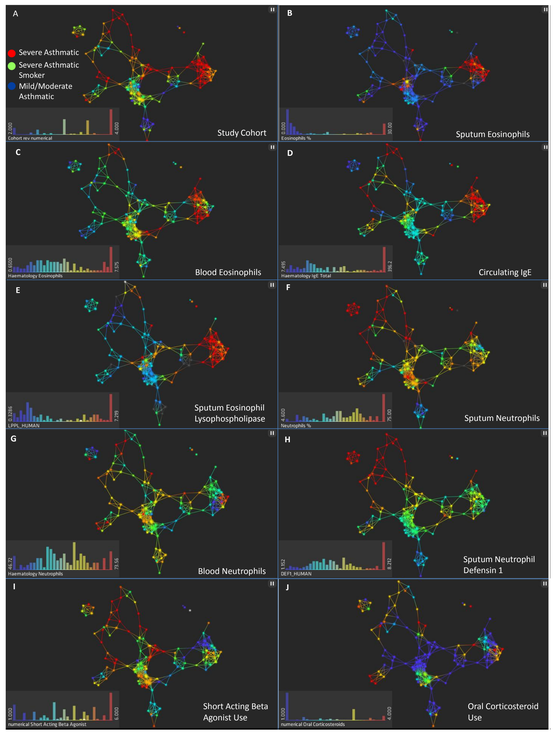


Figure 3. Exploration of the combined graph with clinical and proteomic features. **Panel A** shows distribution of assigned cohorts . **Panels B-E** show various aspects of characterised by eosinophilic inflammation and atopy (sensitisation to common aero-allergens like house dust mites) with subtle differences in airway vs. systemic indicators. **Panels F-H** contrast eosinophilia from previous panels with neutrophilia. Interesting differences in sputum neutrophil counts vs systemic neutrophil counts can also be observed. **Panels I & J** contrast participant use of short acting beta agonists and oral corticosteroids (converted from categorical data). There appears to be correlation between short acting beta agonist use and a subset of participants with sputum neutrophilia

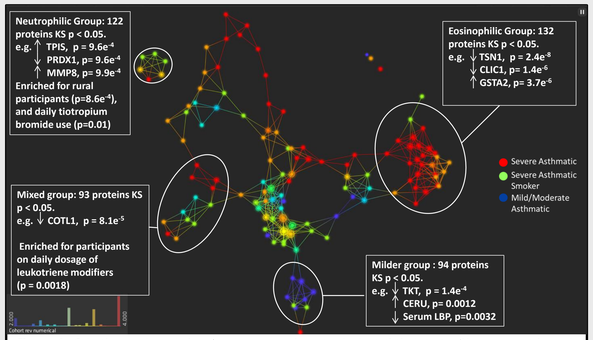


Figure 4. Selection and analysis of subgroups using standard statistical tests (Non parametric and hypergeometric) revealed features of specific groups and their participant membership. This information is being used as a basis for hypothesis generation

Investigating large datasets

Modern high resolution mass spectrometry produces large quantities of spectral data that is often discarded following routine database searching. In order to mine this wealth of sample information we have developed an approach that aligns and groups four dimensional ions generated via high throughput LC-IMS-MS^E (Figure 5). These large matrices, with significant sparsity, are then analysed on the Ayasdi software platform in order to find persistent local groupings of potentially interesting molecular ions (Figure 6) that can then be interrogated for identity and/or fed into classification models.

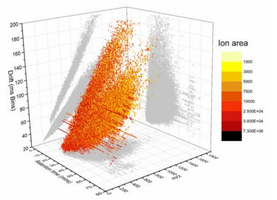


Figure 5. Ion features in a typical sputum LC-IMS-MS^E experiment. The intensities of ~500 000 ions per sample are measured following separation according to their **D**rift time, **R**etention time and **A**ccurate **M**ass (*m/z*), (DRAMI).

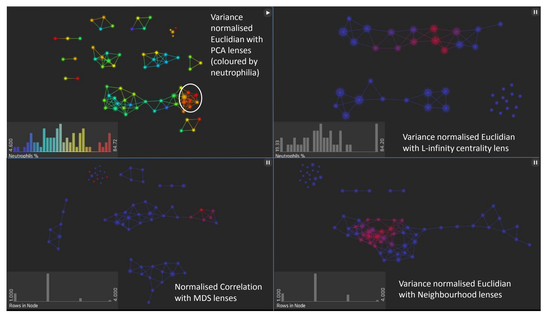


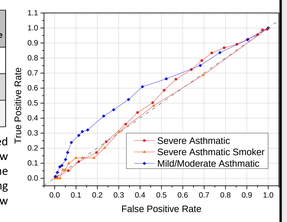
Figure 6. Feature reduced (~70 000 features per sample) TDA graphs generated from DRAMI files were contrasted to clinical information and examined for persistence (red shading) when varying lenses, metrics and analysis resolution. This process is valuable as a feature selection process preceding downstream ion interrogation and classification approaches.

Feature and class selection in cryptic datasets

Traditional statistical analyses of data from high throughput proteomic analysis of serum from asthmatics, produces datasets with very few differentiating features (Table 1 and Figure 6). Using the Ayasdi software platform we were able to identify groups of participants based on these serum data, with specific group assignment guided by persistence of structure and contrasting clinical metadata (Figure 7). These topological groups were subsequently used as class labels in supervised machine learning approaches, implemented in InforSense[4] (e.g. support vector machine, Figures 8 & 9), and resulted in an improved classification performance and predictive models over cohort information alone.

Comparison	Features: p < 0.05		Average Fold Change
	KS test	FDR correction	
Severa Asthma vs. Asthmatic Smokers	3	1	1.18
Severa Asthma vs. Mild/Moderate	15	2	1.31
Asthmatic Smokers vs. Mild/moderate	9	1	1.26

Table 1. Differentially abundant serum proteins assessed using the Kolmogorov-Smirnov test prior to TDA. Very few proteins show differences in abundance levels between the asthmatic cohorts, particularly following multiple testing correction. Those that are statistically different have low magnitude of change. **Figure 7** (right). When cohort assignments were used as class labels in support vector machine analysis, the resulting model displayed poor classification performance



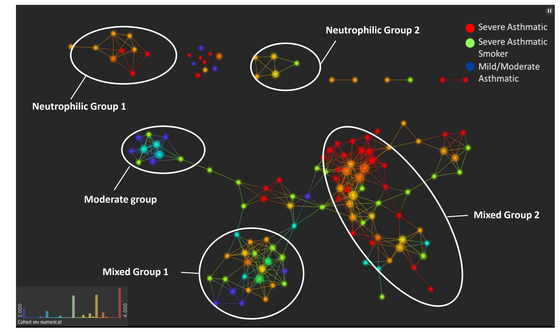


Figure 8. Grapical display of serum data analysed using the Ayasdi platform. Group assignments were based on persistence of groups, and guided by clinical metadata

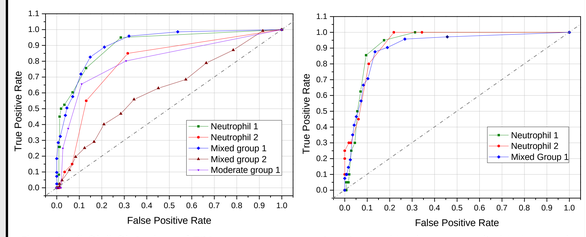


Figure 9. (left) (left) Analysis of TDA groups using support vector machine learning resulted in improved classification models. **Figure 10.** (right) ReliefF[5] and feature evaluation implemented in InforSense software were used to reduce the feature space to 14 proteins that were fed into the classifier. As a result, an excellent model was obtained for three of the class assignments from Ayasdi TDA

References. 1: Lum *et al*., (2013). Extracting insights from the shape of complex data using topology. *Sci reports* **2:** Silva *et al*., (2006). Absolute quantification of proteins by LCMS^E. *Mol Cell Proteomics* **3:** Johnson *et al*., (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* . **4:** IDBS InforSense Suite, http://www.inforsense.com/ **5:** Robnik-Sikonja & Kononenko, (2003). Theoretical and empirical analysis of ReliefF and RReliefF. *Mach Learn*

U-BIOPRED

Innovative Medicines Initiative

efpia European Federation of Pharmaceutical Industries and Associations