study (TB019/NCT1669096), the behavior of a gene signature associated with protection in a controlled human malaria infection study (RTS,S/AS01 vaccine candidate). We indeed hypothesized that this signature would allow us to identify time-points displaying potentially biologically meaningful variation in a vaccine with the same adjuvant, even though protection mechanisms may differ between TB and malaria.

The statistical challenges were to (i) identify clusters of subjects that would present similar biological variation as protected and non-protected subjects from the malaria study (protection status is unknown in the TB study), (ii) validate the findings to control for the bias induced by searching for specific signatures. After signature-driven group assignment and statistical validation using a resampling approach and Monte Carlo simulations, we were able to show the presence of the signature at specific time-points after vaccination in TB019/NCT1669096. The optimal time-points, as defined by the capacity of the signature to reveal response variation, included day 7 post dose 2, which was selected for the ancillary transcriptomics study (C-041-972/NCT02097095).

References

[1] R.A. van den Berg, M. Coccia, W. Ripley Ballou, K.E. Kester, C.F. Ockenhouse, J. Vekemans, E. Jongert, A.M. Didierlaurent, and R. van der Most. Predicting RTS,S vaccine-mediated protection from transcriptomes in a malaria-challenge clinical trial. *Under review*.

Clustering DNA words through distance distributions

A.H. Tavares[a], V. Afreixo[a], P. Brito[b]

[a]CIDMA and iBiMED, University of Aveiro; [b]FEP and LIAAD-INESC TEC, University of Porto

Session: Biomedics, Room: EA3 *Wednesday 12th, 12:00 – 12:20*

Functional data appear in several domains of science, for example, in biomedical, meteorologic or engineering studies. A functional observation can exhibit an atypical behaviour during a short or a large part of the domain and this may be due to magnitude or to shape features. Over the last ten years many outlier detection methods have been proposed. In this work we use the functional data framework to investigate the existence of DNA words with outlying distance distribution, which may be related with biological motifs.

A DNA word is a sequence defined in the genome alphabet $\{ACGT\}$. Distances between successive occurrences of the same word allow defining the *inter-word distance* distribution, interpretable as a discrete function. Each word length k is associated with a functional dataset formed by 4^k distance distributions. As the word length increases, greater is the diversity of observed patterns in the functional dataset and larger is the number of distributions displaying strong peaks of frequency.

We propose a two-step procedure to detect words with an outlying pattern of distances: first, the functions are clustered according to their global trend; then, an outlier detection method is applied within each cluster. Each distribution trend is obtained by data smoothing, which avoids some distributions' peaks, and similarities between smoothed data are explored through hierarchical complete linkage clustering. The dissimilarity between functions is evaluated using the Euclidean distance or the Generalized Minimum distance [1], which considers the dependence between domain points. The resulting dendograms are then cut leading to a partition of the distance distributions. For the second step we use the Directional Outlyingness measure which assigns a robust measure of outlyingness to each domain point and is the building block of a graphical tool for visualization of the centrality of the curves [2].

We focus on the human genome and words of length $k \leq 7$. Results are compared with those obtained by applying only the second step of the procedure [3].

Keywords: distance distribution, DNA word, directional outlyingness.

References

[1] X. Zhao, E. Valen, B.J. Parker, and A. Sandelin (2011). Systematic clustering of transcription start site landscapes. *PloS one*, **6**(8), e23409.

[2] P.J. Rousseeuw, J. Raymaekers, and M. Hubert (2016). A measure of directional outlyingness with applications to image data and video. arXiv:1608.05012

[3] A.H. Tavares, V. Afreixo, P. Brito, and P. Filzmoser (2016). Directional outlyingness applied to distances between genomic words. In *RECPAD 2016*, 108–110.

Lifting and clustering

N. Bozkus

University of Leeds

Session: Clustering, Room: EA4 *Wednesday 12th, 11:00 – 11:20*

A popular question in hierarchical clustering is how many clusters exist in a data set, or where to 'cut the tree'. Even though many methods have been proposed, this topic still attracts the interest of researchers. Previous indices capture the number of clusters quite well if clusters are well separated, but when the clusters overlap or have unusual shapes, their performance deteriorates. I propose a new method based on a multiscale technique called lifting which has recently been developed to extend the 'denoising' abilities of wavelets to data on irregular structures.

My method applies lifting to the structure of a dendrogram. I then assume that the distances between data points and cluster centroids are affected by noise. Denoising the mean distances of data points from each cluster centroid helps decide where to cut the tree. This method will be illustrated with both real and simulated examples.

Keywords: wavelets, lifting, cluster validity index.