H36: Pooya Zakeri, Leon-Charles Tranchevent and Yves Moreau. **Kernel-Based Gene Prioritization Using Late Integration versus Geometric Kernel Fusion**

Abstract: In biology, there is often the need to discover the most promising genes among large list of candidate genes to further investigate. In the recent years, several computational approaches based on different genomic data sources and often machine learning methods have been used to crack this problem efficiently. While a single data source might not be effective enough, fusing several complementary genomic data source results in more accurate prediction. Finding an efficient and cost-effective technique for merging these complementary genomic data has received increasing attention.
We consider several genomic data sources including annotation-based data sources, expression data sources, protein-protein interaction networks and literature, as well as sequence-based protein features such as information extracted directly from position-specific scoring matrices and local sequence alignment.
Integrating data sources at the decision level, such as in the ensemble learning framework, is considered as an intuitive manner to deal with heterogeneous data sources. We discuss the advantage of modeling multi-heterogeneous biological data fusion in the prioritization task based on the ordered weighting averaging (OWA) operator and robust rank aggregation (RRA). We design several kernel-based gene prioritization frameworks that integrate multiple genomic data sources through late integration. After transforming genomic data matrices into kernel matrices, one kernel-based ranking approach is trained with known disease genes on each data type. Then fusion strategies defined by the use of OWA operator weights and ordered statistics weights are employed to aggregate performance rankings.
Moreover, we propose a kernel-based gene prioritization framework through genomic data fusion which is our kernel-based data fusion framework, recently developed and considered as a powerful tool for protein fold classification. It has been observed that geometric data fusion is less sensitive in dealing with complementary and noisy kernel matrices compared to typical multiple kernel learning approaches. Since genomic kernels often encodes the complementary characteristics of biological data, this motivates us to see the application of geometric data fusion in the gene prioritization task.
Finally, we evaluate our models by applying them to the disease data sets on which Endeavour has been benchmarked. Our late integration frameworks based on OWA or RRA offers an improvement in empirical performance. In addition, it is observed that using geometric kernel fusion, we achieve competitive results compared with best existing MKL approaches proposed for gene prioritization task.