MACHINE LEARNING AND GRAPH METHODS TO STUDY AGING AND EPILEPSY

PhD thesis defense, handout

Tibor Nánási MD

Semmelweis University

János Szentágothai Doctoral School of Neurosciences

Supervisor: István Ulbert, MD, DSc

Budapest
2020

Introduction

Support Vector Machines (SVM), as well as graph-based data analysis can be effectively used in various fields of neuroscience, including neurophysiology and molecular biology. Previous works of the author include applications of graph theory on molecular (Simkó et al. 2009; Kiss et al. 2009) as well as neural systems (Nánási et al. 2016; File & Nánási et al. 2020). Statistical (File & Nánási et al. 2020; Banlaki et al. 2015; Kovacs-Nagy et al. 2013) and model fitting tools (Lehallier et al. 2019) have been employed extensively in another studies.

In this Thesis, we will explore the synergistic application of machine learning and graph techniques to study large-scale functional brain activity of epileptic patients using semi-invasive electrocorticography (ECoG) as well as transcriptomic and proteomic changes relatable to organism-level and brain aging. Also, viability of parabiosis to study age-related alternations will be investigated using novel bioinformatic methods.

Objectives

In pharmacologically intractable epilepsy, Seizure Onset Zone (SOZ) detection is a crucial diagnostic step needed to plan curative surgery. To gain insights on key epileptic features of the ECoG signal, various SVM based models will be evaluated. Feasibility of machine learning tools to reproduce expert diagnosis will be investigated.

In the fields of transcriptomics and proteomics, the number of integrative studies involving both graph and machine learning techniques is limited when compared to the extensive literature on both topics. Aiming to analyze omics measurement in context of a priori knowledge on gene product interactions, a novel integrative method, the Predictome approach will be introduced and validated on repeated transcriptomic measurements of human brain aging. Finally, proteomic changes elicited by heterochronic parabiosis will be explored using the validated model to ponder on the relevance of parabiosis and the importance of evolutionally conserved, circulating factors in blood plasma in the context of brain aging.

Methods

Electrocorticography

Deep sleep recordings free from obvious epileptic activity from six patients suffering from pharmacologically intractable focal epilepsy have been acquired and processed as described in preceding works (Nánási et al. 2016; File & Nánási et al. 2020). ECoG implantation and monitoring was performed in the Department of Functional Neurosurgery and Center of Neuromodulation, National Institute of Clinical Neurosciences, Budapest, Hungary. Seizure Onset Zones were identified by experts in the Epilepsy Centrum, Department of Neurology of the same institute.

Classification of ECoG recordings using SVM

Based on their channel of origin, epochs were assigned to SOZ and non-SOZ classes and related frequency band specific amplitude information was presented to the machine learning algorithm (SVM) as observations. The

models were tasked to reconstruct expert opinion on SOZ or non-SOZ status. 5-fold cross-validation was used during training and evaluation. Kernel choice effects were investigated by comparing the output of linear and Gaussian (RBF) kernels using standardized predictors and default kernel parameters. Performance was measured as Matthews Correlation Coefficient (MCC).

Transcriptomics and proteomics

To study human brain aging, genome-wide RNA sequencing data obtained from the GTEx Consortium was analyzed. Notably, the dataset contains a rather unique redundancy for frontal cortex and cerebellum as in a considerable subset of cases tissue samples from the exact same individuals were re-sampled and measured by two independent laboratories. This feature was used for testing and validation purposes. It has been shown recently (Lehallier et al. 2019) that aging can be characterized by multiple, temporally separated waves of proteomic changes in the blood plasma. Crests of the undulating proteome alternations are reached at ages of 34, 60 and 78

years of age and they are reflecting distinct biological pathways, implying profound temporal heterogeneity of the aging process. The 1st and 2nd waves were marked as points of interest for our transcriptome analysis using the GTEx data.

Parabiosis is a surgically induced state that connects the circulatory systems of multiple organisms. Experiments showed that multiple tissues, including brain, can be functionally and structurally rejuvenated in old mice connected to young (heterochronic parabiosis). Also, aging phenotype can be provoked in the opposite direction. Here, blood plasma proteomic changes in such states were investigated using previously published measurements (Lehallier et al. 2019) employing SomaLogic aptamer technology.

Interactome (PPI) data

To obtain a deposit of biologically meaningful Protein-Protein Interactions (PPI) the OmniPath and PICKLE databases were merged. This produced a novel set of

highly dependable PPI library describing 195.456 interactions between 16.005 proteins which is, to our knowledge, could be regarded as the most comprehensive body of reliability-optimized interactome information currently available. Here, UniProt namespace was used to avoid mapping artefacts.

Predictome: integrating genome-wide molecular measurements with interactome information

Each link (interaction) of the PPI network was considered as a separate model fitting task to be solved. Two continuous predictors were defined as the gene expression levels or direct concentration readouts corresponding to the interactors. Models were trained to classify the samples into younger-older or iso- and heterochronic parabiotic groups in a binary manner using repeated cross-validation. Performance metrics were assigned to each link to define its weight. This process transforms the nascent PPI into a phenotype-specific weighted network termed as the “Predictome”.

Gene ranking and Functional Enrichment Analysis

To evaluate the importance of individual genes, information represented by the Predictome graph must be projected to node level. Eigenvector Centrality (EC) has been shown to be informative in various biological problems by emphasizing elements of super-regulatory role or critical targets of regulatory pathways. Here, it highlights genes whose expression levels are correlated to phenotype or contributors of interactions where the combined information could be used as a successful predictor of it. To compensate for the topological bias arising from the sparsity of the Interactome, surrogate EC values were calculated in 100.000 networks with shuffled link weights. This way, Monte Carlo probability of measured EC can be assessed, on which basis the genes could be ranked. Ranked lists were subjected to Sliding Enrichment Pathway Analysis (SEPA, Lehallier et al. 2019) using the Reactome pathway database. To quantify model stability across replicated laboratory measurements, Jaccard Similarity Index was calculated.

Results

SOZ detection requires complex spectral information

Relying on single-band power features, regardless of the used frequency band or applied kernel transformations, SVM was unable to find separation planes reliably sorting epoch data to SOZ or non-SOZ origin. In contrast, SVM models outfitted with full spectral information provided better results for both linear and Gaussian approaches. Importantly, Gaussian kernel SVMs were able to reproduce expert decisions with great fidelity in all patients.

SVM unmasks information embedded in gene interactions

In multiple occasions link based SVM models could predict phenotype even if the linked genes did not show altered expression when analyzed separately. This demonstrates the ability of machine learning algorithms to extract hidden information inaccessible to methods operating only on singular gene expression levels. For example, expression of microtubular genes TUBB4B and

TUBG1 did not change during aging but their relative concentration shifted towards TUBB4B dominance. Complementary information of connected genes improved age prediction for more than 43.000 interactions.

Genes influenced by Brain Aging and Parabiosis are functionally connected and validated by the literature

Extensive interactome areas found to be associated with aging, covering 8.82-15.81% of the measured genome. Genes associated in both repeated measurements with at least 2 out of the 4 studied aging phenotypes (2 aging waves in 2 brain areas) were selected and combined with elements influenced by Parabiosis (either by provoked aging in young animals or rejuvenation in old animals). The set found to be enriched for PPI interactions ($p < 10^{-16}$) confirming the existence of a dense aging subnetwork. Also, predicted aging genes are strongly enriched in elements curated in the GenAge aging database ($p_{Fisher} = 3.3*10^{-25}$) and frequently mentioned together in PubMed abstracts with keywords *"Aging"* and *"Age-related"* ($p_{Wilcoxon} = 1.5*10^{-78} – 8.7*10^{-67}$).

Interactome information enhances abundance and reproducibility of Pathway Analysis results

Inclusion PPI information resulted in enhanced interpretability in context of a priori biological knowledge, highlighting 6.39-17.26% of the Reactome namespace. This was a marked improvement over a control analysis disregarding PPI network information (0-2.1%). Comparing results arising from replicated laboratory measurements, robustly enriched pathways were virtually absent with single gene analysis. In contrast, Predictome-driven results showed more consistency with a Jaccard Index of 0.38 across enriched pathway sets.

Brain Aging and Parabiosis sharing functional aspects

Pathway associations with confirmed significance showed considerable heterogeneity across brain areas and aging waves. In parabiosis, provoked aging had more profound functional impact compared to rejuvenation (131 and 29 pathways). These findings agree with previous literature. Also, the 271 pathways linked to multiple aging

phenotypes or to parabiosis were extensively referenced in PubMed abstracts on age-related conditions ($p_{Wilcoxon} = 6.7*10^{-5}$).

Five biological processes, including *altered regulation through small ubiquitin-like modifier (SUMO)* proteins found to be mutual aspects of all aging phenotypes and the same time influenced by parabiosis. *Splicing of mRNA* seems to be altered exclusively in early aging whereas late aging could be characterized by 108 pathways involving *axon guidance, hemostasis, integration of energy metabolism and innate immune functionality*.

In parabiosis, accelerated aging effects elicited by old blood resembled the patterns found in early neocortical aging. Changes related to *apoptosis* and *cellular senescence* were common, together with altered signaling through *RUNX2*, *PTK6* and *TGF-beta*. Rejuvenation effects were relatively isolated but relatable to *EGFR signaling*, *Clathrin-mediated endocytosis*, and to the *complement system*. Furthermore, involvement of *ERBB2*, *ERBB4*, and *PTK6* pathways was predicted.

Discussion

Electrophysiology of epilepsy

Previous publications of the author shown the usability of network-based models (File & Nánási et al. 2020) and multi-modal feature integration (Nánási et al. 2016) in SOZ localization from ECoG recordings. Here, the feasibility to deduct diagnostic information from resting state, seizure-free neural activity was demonstrated. The uniform, stereotypical nature of non-REM phases was successfully exploited to reduce the potential effects of sensory stimuli or altering mind states.

SOZ detection accuracy achieved using heuristic methods (Nánási et al. 2016) was matched using general purpose machine learning tools. However, full spectral information and Gaussian transformation of the feature space was needed to reproduce expert decisions, highlighting the complexity of the problem. Epileptic patterns were found to be patient-specific and therefore, much bigger repositories of ECoG data would be needed to construct a generalizable diagnostic tool.

The concept of the Predictome

To study aging in the molecular level, we applied a novel integrative approach to investigate transcriptomic changes related to two stages of human neocortical aging. Furthermore, we compared our findings with the effects of heterochronic murine parabiosis on the plasma proteome. Our method is based on the analysis of the Predictome, which is defined as an Interactome (PPI) network weighted according to the predictive performance of SVM models fitted to gene expression data or direct plasma protein concentration measurements corresponding to each link. The significance of Eigenvector Centrality (EC) of elements in this weighted graph was quantified using Monte Carlo statistics, summarizing both local and global topological characteristics of the resulting network on a node (protein) level.

Graph theory offers an especially well adaptable platform for integrating transcriptomic or proteomic readouts with curated protein-protein interaction information. As the used facet of a priori knowledge describes the elementary

building blocks (interactions) of the biological system, it is less biased by constantly evolving conceptual categories than higher level representations (like pathways). The Predictome representation also helps to deal with genes whose role could only be assumable in context. Additionally, the used EC metric incorporates non-local features of the graph when quantifying gene importance.

Multiple other approaches are present in the literature to exploit a priori knowledge encoded in the Interactome when analyzing high-throughput omics data. These methods are, usually, building functional maps of systems corresponding to separate phenotypes by modifying or thresholding the PPI network based on co-expression measurements, then comparing the results of secondary measurements on these maps to deduct conclusions. The Predictome approach is fundamentally different from such techniques as instead of relying post-hoc comparisons, the relevance of interactions regarding to phenotype is encoded into the model implicitly.

Validation of the Predictome approach

Aging can be characterized by widespread, but subtle changes of transcriptomic activity. Because of the marked disproportionateness of the variable space (number of genes) and the sample size (number of subjects), false positive results can be expected, and robustness becomes a key feature of the analytic pipeline. As demonstrated using repeated measurements on the experimental level, the Predictome approach surpassed the single-gene technique both in abundance and in reproducibility of Reactome term enrichment results.

Conclusions were highly coherent with literature. Both on gene and pathway level, the highlighted items were frequently referred together in PubMed abstracts with *"Age-related"* alternations or *"Aging"*. Furthermore, genes enlisted in the GenAge aging database were greatly enriched among critical elements of aging Predictomes.

Implications on the biology of brain aging and parabiosis

Various processes related to immune- and vascular functions, development, survival, senescence, and cellular death were highlighted in aging Predictomes. *Toll-like receptor cascades, Interleukins, VEGF, TGF-β* found to be instrumental in both aging waves, in agreement with the literature. Also, novel predictions emerged: involvement of *SCF-KIT, NTRK1, ERBB2, ERBB4,* and *PTK6* signaling, together with *Sema4D, EPHA, and EPHB* were implicated only indirectly in aging and neurodegeneration so far.

Previously established findings implying the dominance of non-linear changes in the aging plasma proteome (Lehallier et al. 2019) were found to be valid in case of aging brain transcriptomes as well, emphasizing the interplay across tissue aging and circulating factors. Furthermore, the relevance of parabiotic models to study human brain aging could be verified. Young parabionts exposed to old plasma, a phenotype relatable to accelerated aging shared multiple aspects with human

brain aging. These features included alternations of *Cell Cycle* mechanisms, *DNA Repair* and *SUMOylation*. Rejuvenation effect elicited by young plasma on old parabionts is known to be more limited which was well reflected by the confined set of associable Reactome terms. However, similarities could be found with both stages of human brain aging including changes in *Metabolism of proteins* and *ERBB2* signaling (early-stage) or *PTK6* and *ERBB4* signaling (late-stage) in a robust manner. Together with recent findings on the intervened nature of systemic milieu and brain aging, those results confirm the importance of plasma measurements and parabiosis to investigate brain aging and related diseases.

Publications of the author relevant to this thesis

File, Bálint*, **Tibor Nánási***, Emília Tóth, Virág Bokodi, Brigitta Tóth, Boglárka Hajnal, Zsófia Kardos, et al. 2020. "Reorganization of Large-Scale Functional Networks during Low-Frequency Electrical Stimulation of the Cortical Surface." ***International Journal of Neural Systems*** 30 (3): 1–15. https://doi.org/10.1142/S0129065719500229.
**** contributed equally***

Nánási, Tibor, Bálint File, Emília Tóth, László Entz, István Ulbert, Dániel Fabó, and Lóránd ErÅ'Ss. 2016. "Synergism of Spectral and Coupling Modalities in Epileptic Focus Localization from IEEG Recordings." ***PRNI 2016 - 6th International Workshop on Pattern Recognition in Neuroimaging***, 2–5. https://doi.org/10.1109/PRNI.2016.7552354.

Lehallier, Benoit, David Gate, Nicholas Schaum, **Tibor Nanasi**, Song Eun Lee, Hanadie Yousef, Patricia Moran Losada, et al. 2019. "Undulating Changes in Human Plasma Proteome Profiles across the Lifespan." ***Nature Medicine***. https://doi.org/10.1038/s41591-019-0673-2.

Lehallier, Benoit, David Gate, Nicholas Schaum, Tibor Nanasi, Song Eun Lee, Hanadie Yousef, Patricia Moran Losada, Daniela Berdnik, Andreas Keller, Joe Verghese, Sanish Sathyan, Claudio Franceschi, Sofiya Milman, Nir Barzilai, Tony Wyss-Coray. 2019. “Undulating changes in human plasma proteome across lifespan are linked to disease.” ***bioRxiv***. https://doi.org/10.1101/751115