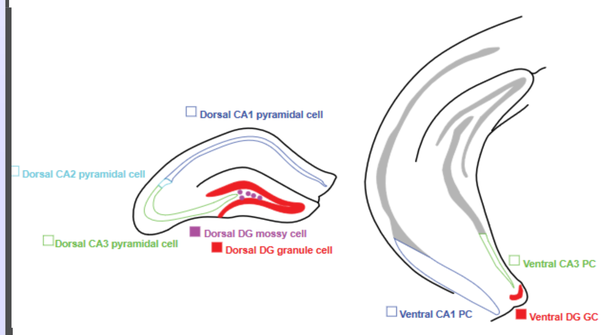
Granule and mossy cell gene expression study: single cell sequence analysis and ring monitored by EEG with Bayes Variational Inference

Tampereen yliopisto
Tampereen ammattikorkeakoulu

Qin He, Jarri Hyttinen, Sampsa Pursiainen
qin.he@tut.fi

Tays
Elämän tähden

Abstract



The two typical excitatory cell types granule cell(GCs) and mossy cells of DG in chronically epileptic mice identified by optogenetic tagging are studied through the electrographic data of the seizure.To investigate the consequences of MC and GC loss on cognition, the activities of the cell neurons are recored by MEA Further, gene expression profiling is applied to understand the functionality of them in addition to the intrinsic level. The major excitatorial neuronal classes of the hippocampus(figure beyond), GC, and MC of the dentate gyrus reveals some unknown properties to the epilepsy. The mouse model with hippocampal sclerosis, uses a combination of optogenetic, electrophysiologycal, and behavioral approaches connecting the light-response features to gene expression in mammalian cells. One machine learning based on Bayes Variational Inference model is also applied to the t-SNE clustering of the genes. As DNA and RNA are usually of high dimension, the Bayes variance is utilised to modify the BIC in guaranteeing the convergence, computing KL and perplex and the cluster distance are measured with Euclid metric. One review comparing epilepsy with AD in genetic regulation is introduced as well for giving guide on epilepsy gene detection method and pathology study at hippo-campus region comparing epilepsy and AD.

Model Formulation

The Bayes Variational Inference we use here is to minimize the Kullback-Leiber(KL) divergence, the log difference between observed and approximated posterior which can also be regarded as an optimization problem: Modifying the t-Distributed Stochastic Neighbor Embedding(t-SNE), the pairwised distances in high dimensional space with data points x is converted to corresponding embedding points y pairwise join distributions in low dimensional spaces, which respectively follows:

$$q_{i,j} = \frac{(1+|y_i - y_j|^2)^{-1}}{\sum (1+|y_s - y_t|^2)^{-1}}$$

while high dimensional one is defined in symmetrical conditions:

$$p_{i,j} = (p_{i|j} + p_{j|i})/2n)$$

,where

$$p_{i|j} = \frac{exp(-|x_i - x_j|^2/2\sigma_j^2)}{\sum exp(-|x_s - x_t|_{2/2\sigma j^2}))}$$

and the KL to be optimised is thus:

$$KL(P||Q) = \sum p_{i,j} * log \frac{p_{i,j}}{q_{i,j}}$$

Note that the σ is optimized through bisectional search automatically with the pre-specified perplexity j

$$Perplex(p_j) = 2^{H(p_j)}, \text{ where } H(P_j) = -\sigma_j p_{i|j} * log^{i|j}$$

so that

$$Perplex(p_j) = Perplex$$

, where Perplex is the hyperparameter of the t-SNE central to the final cluster.
Large Perplex usually leads to the embedding suboptimal in detecting the pattern of the data(In the limit, when the Perplex goes to the number of data points, the corresponding embedding form a Gaussian or uniform like distribution and fails to be useful for structure detection at all) and thus, we design a new criteria:

$$S(Perplex) = KL(P||Q) + log(n) * \frac{Perplex}{n}$$

with inspiration of the Bayesian Information Criteria(BIC):

$$BIC = -2 * log(L) + log(n) * k$$

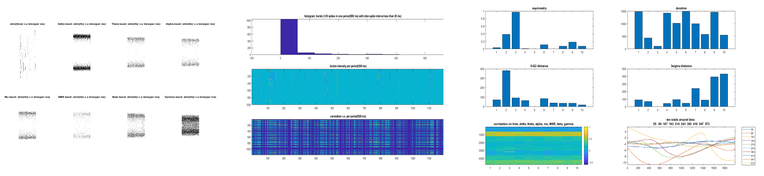
, where the first term stands for the goodness-of-fit of the maximum-likelihood-estimation and the second controls the complexity of the model with penalty k scaled by log(n).

Performance in Clustering with Bayes Variational Inferenced t-SNE

Intuitively, when Perplex increases, differences among points will become less and less significant with regard to the length of the kernel in distribution P, and P will tend to uniform.The forward form of KL has large cost for under-estimating probability but not for over-estimating. That is, if $p_{i,j}$is large and $q_{i,j}$ is small, KL divergence is large while in the opposite direction, KL is not affected. Increasing Perplex leads to larger σ and more uniform $p_{i,j}$ so it is easier is for the student-t distribution in low dimensional space to assign mass for all probability points sufficiently. This is the so called crowding problem: When projecting from high to low dimensional space, there is not enough room in lower dimensional space.Generally, increasing Perplex relaxes the problem and reduces the amount of structure to be modelled with less error according to KL while pays a cost in the second term. With the practical test, we apply it on the MC GC cell expression classification with t-SNE.

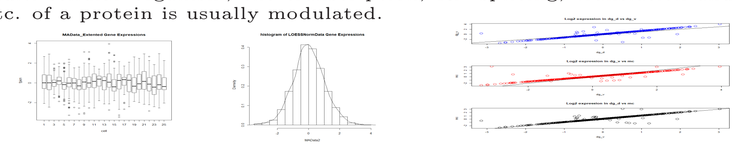
MEA Analysis

There are 5 ictal periods recorded (10 bins of interested separately.). For the whole dataset, the auto-correlation is calculated to detect the onset of seizure and preprocess including baseline correlation and normalization and filtering are applied. As the seizure is usually recorded on high frequency band compared to other signals, for instance the signal for learning tasks, the MEA signal is studied comparing different bands as well as on the whole 180 s time period. According to the middle figure, the signal is oscillated evenly on the gamma band (25 Hz to 100 Hz) and not detected as spikes at delta, theta, alpha, mu, SMR and beta bands according to the fpkm values (on log level) of the genes in the range -3 to 4. Specifically, bursts(defined as more than 30 spikes detected in one period: 500 ms with inter-spike interval less tan 20 ms) are detected mainly in the time period: 0-80ms and is sparsely detected at 80ms to 410 ms(in the first 500 ms around ictal 1). The intensity is increased at 100-110 ms, 460-480 ms , 790-810 ms and 1100-1200 ms significantly without significant correlation. (See rightest figure)
As the time window shows negative correlation (Pearson) round 100 ms, it is not significantly correlated on any band with only a little bit synchronization on delta and gamma band (left bottom figure). And finally, the 10 bins are classified into narrow and broad spikes according to its asymmetry, duration, quantile and deviation distribution. (rightest figue) As the majority of the spikes are biased, they are classified as broad type.[1]

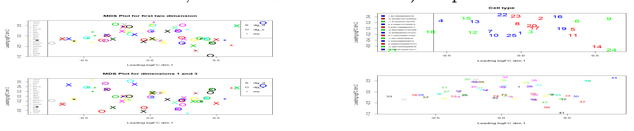


Genetic analysis

As study of cellular processes existing inside intact organisms requires methods to visualize cellular functions such as gene expression in deep tissues, the 75 expression dataset (dg-d, dg-v and MC cell separately according to left figure) is studied along with the gene regulation as the substrate of the observable trait at which the genotype gives rise to the phenotype (further cellular differentiation adaptability of organism in the versatility and adaptability). After the LOESS normalization, it is normalized according to QC test(second figure) and most genes of mc and dg-d are expressed stationarily with MA value mainly distributed around zero (since the expression of genes are assumed to be unchanged at certain tissue region.) while has decreased trend for the dg-v expression.(third figure). Detailed, the dg-v to dg-d ratio is shown with higher skewness in the top figure while with skewness around 0 of mc to dg-v and mc to dg-d in the two bottom figures. The 13, 16 and 18 numbered genes are relatively densely expressed. As the profound effect on the funcitions (actions) of the gene in a cell is usually studied as a multi-cellular organism, the transcription, RN splicing, translation and etc. of a protein is usually modulated.

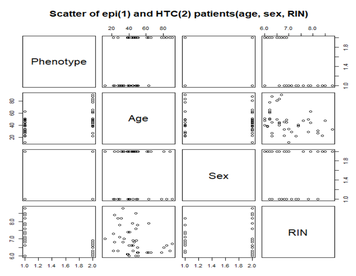


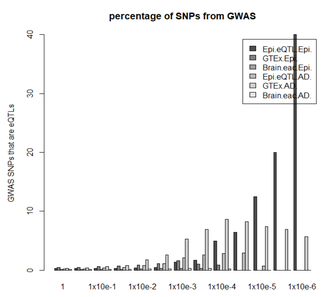
The data is 25 dimentioned with some specific biomarkers (For instances, the GTpases related Cdc45). Thus, the MDS plots for the first, second and third can be seen here: bottom. The cell with different markers can be seen with their different expression levels. As the 13 and 18 marked gene, Prox1 and St18 (red and green) are expressed with larger density in dimension 2 than dimension 1, where the marcks1 (yellow) are contrarily distributed evenly of different levels on both dimension 1 and dimension 2(along with extended dataset 58 53 74 26 18 34 66 and 42, 50 19 11 67 and 43) expressed more in cell



Review related to molecular analysis studying epilepsy and AD pathology

Athough our study is based on mice hippocampus data, there is always the aim to analogue some similar pathophysiology on human epilepsy. The molecular genetics is well used on the prominent study especially combining cognitive and behavior autosomal dominant forms of epilepsy, genetic linkage studies, positional cloning, (GWAS)genome wide association studies and (eQTL) expression quantitative loci analysis. As seizure is also commonly detected along with (AD)Alzheimer's disease, here in addition to the disease-associated variants from an epilepsy GWAS, the meta-analysis of AD associated with AD is also reviewed. As left figure shows, epilepsy patients are with higher RIN($p = 3.45185e-7$). As idealized EEG trace would have an RIN approximately at 10, the epilepsy patients have the averagely lower RIN. Intuitively, the unlike HC group, epilepsy group's RNAs are usually shrunk giving samples at most left and closer to where marker originally was. Further more, the average age level of the epilepsy group is also higher than the HC group($p = 0.001049$). Among the epilepsy group, the number of those with earlier onset is more than those with late onset(opposite to the AD patients according to [2]). Meanwhile, sex is not a significantly different factor between epilepsy group and HC group($p = 0.062437$). In the middle figure, the significant SNPs of epi-eQTLs folding differences fraction to all increases soarly with the decreases of epilepsy GWAS meta-analysis p-value. And the speed increases with the (fdr)false detected rate decrese as well. However, the three fdr correction does not give significant differences.($p = 0.05741$, 0.05, 0.0654) In the right figure, there lists the epi-eQTLs enrichment analysis within Mendelian eilepsy genes.Mendelian epilepsy syndromes: Mendelian disorders in which epilepsy or seizures are a cardinal feature. Data from the Online Mendelian Inheritance in Man (OMIM) database.





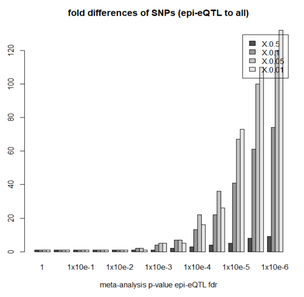
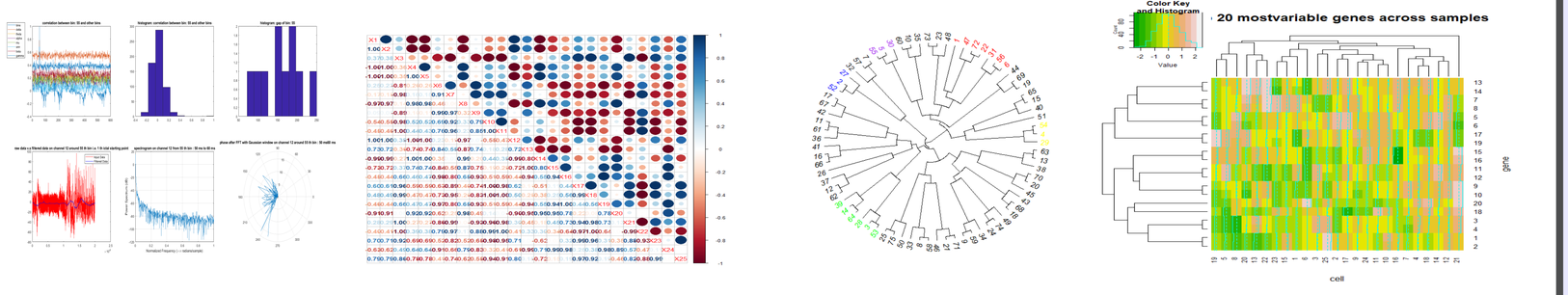


Table S8. Mendelian epilepsy syndromes. Mendelian disorders in which epilepsy or seizures are a cardinal feature. Data from the Online Mendelian Inheritance in Man (OMIM) database.

Gene	Syndrome
AARS	Epileptic encephalopathy, early infantile, familial adult, 2
ADGRV1	Febrile seizures, familial, 29
ADRA2B	Epilepsy, myoclonic, 4,
ALDH7A1	Epilepsy, pyridoxine-dependent
ALG13	Epileptic encephalopathy, early infantile, 36
ARHGEF9	Epileptic encephalopathy, early infantile, 8
ARX1	Epileptic encephalopathy, early infantile, 38
ARX2	Epileptic encephalopathy, early infantile, 1
ASAH1	Spinal muscular atrophy with progressive myoclonic epilepsy
BFIS1	Seizures, benign familial infantile, 1
BFIS4	Seizures, benign familial infantile, 4
BRAT1	Rigidity and multifocal seizure syndrome, lethal neonatal
CACNA1D	Primary aldosteronism, seizures, and neurologic abnormalities
CACNA1H	Epilepsy, childhood absence, susceptibility to, 6
CACNA1H	Epilepsy, idiopathic generalized, susceptibility to, 6
CACNB4	Epilepsy, juvenile myoclonic, susceptibility to, 6
CACNB4	Epilepsy, idiopathic generalized, susceptibility to, 9
CASR	Epilepsy idiopathic generalized, susceptibility to, 8
CDKL5	Epileptic encephalopathy, early infantile, 2
CERS1	Epilepsy, progressive myoclonus, 8
CHD2	Epileptic encephalopathy, childhood-onset
CHRNA2	Epilepsy, nocturnal frontal lobe, type 4
CHRNA4	Epilepsy, nocturnal frontal lobe, 1
CHRNB2	Epilepsy, nocturnal frontal lobe, 3
CLCN2	Epilepsy, idiopathic generalized, susceptibility to, 11
CLCN3	Epilepsy, juvenile absence, susceptibility to, 2
CLCN2	Epilepsy, juvenile myoclonic, susceptibility to, 8
CLN8	Ceroid lipofuscinosis, neuronal, 8, Northern epilepsy variant
CNNM2	Hypomagnesaemia, seizures, and mental retardation
CNTN2	Epilepsy, myoclonic, familial adult, 5
CNTNAP2	Cortical dysplasia-focal epilepsy syndrome
CPA6	Epilepsy, familial temporal lobe, 5
CPA6	Febrile seizures, familial, 11
CSTB	Epilepsy, progressive myoclonic 1A
DEPDC1	Epilepsy, familial focal, with variable foci
DIAPH1	Seizures, cortical blindness, microcephaly syndrome
DNM1	Epileptic encephalopathy, early infantile, 31
DOCK7	Epileptic encephalopathy, early infantile, 23
ECA1	Epilepsy, childhood absence, 1
ECT	Centrotemporal epilepsy
EEF1A2	Epileptic encephalopathy, early infantile, 33
EFHC1	Epilepsy, juvenile absence, susceptibility to, 1

Conclusions



In summary, according to left figure the correlation is largest within one window on delta band and gamma band (with the maximum around 90 ms. The histogram and the gap can be seen distributed with gamma and normal distribution. According to the filtered signal, the type of the spike is labeled with asymmetricity (most weighted factor), quantile, duration of the spike which is related more to the time information.(The classification results is not with high enough accuracy mainly because the 10 ictal signal bins after the ANOVA is not distributed normally and will be modified in the next report.) However, the spectrogram and polar gram shows the information on phase domain which reveals the phase-lock property of the electrode signal also supporting the truth that the seizure component of the signal fires at the higher frequency. With the normalized expression data ($p = 0.0035$ with ks-test), Genes expressed in the 3-sigma distance, Cdc45, Prox1, St18, Marsckl1, SOX2, Cdk4 and etc. known to have distinct functions are shown in the heatmap and accordingly posses correlation with close to zero in the correlation figure(second figure). However, to explore the difference of the function on cellular scale, the welch two sampled t-test is applied. According to the result, the gene expression for dg-d, dg-v and mc($p = 0.009724$, $p = 0.008046$ and $p = 0.008584$) are not of the same mean on log2 level and thus clustered in the third figure. Generally, dg-d and mc are expressed more evenly while dg-v is distributed on some specific dimeension (correction method will be applied in the next step exploration. Top 20 expressed genes can be found in the rightest figure.) Further more, there can be further exploration on the cell difference and evolutaion with both time and structure information.(TF will be analyzed with pseudo-time in the next step.There is the trial of the pseudo-time analysis shown in the first class of esc with first 70 dataset distributed with the coordinate(2D) recorded by the prone and expression in the microarraydata.) To explore the transcription of DNA in the future, we also introduced one review of genetic regulation of gene expression in human hippo-campus.