Work report on RNA-seq gene expression of hippocampal principal neurons

qin.he

October 2019

1 Introduction

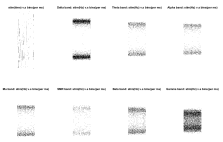
The two typical excitatory cell types granule cell(GCs) and mossy cells of DG in chronically epileptic mice identified by optogenetic tagging are studied through the electrographic data of the seizure. To investigate the consequences of MC and GC loss on cognition, the activities of the cell neurons 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, GC, and MC of the dentate gyrus reveals some unknown properties to the epilepsy. The mouse model with hippocampal sclerosis, uses a combintion of optogenetic, electrophysiologycal, and behavioral approaches connecting the light-response features to gene expression in mammalian cells.

2 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 1 , the signal is oscilated 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 2) As the time window shows negative correlated (Pearson) round 100 ms, it is not significantly correlated on any band with only a little bit synchronization on delta and gamma band (3). And finally, the 10 bins are classified into narrow

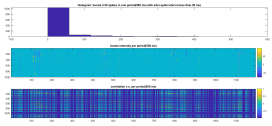
该文档是极速PDF编辑器生成，
如果想去掉该提示,请访问并下载：
http://www.jisupdfeditor.com/

and broad spikes according to its asymmetry, duration, quantile and deviation distribution. (4) As the majority of the spikes are biased, they are classified as broad type.



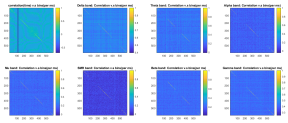
and f).png

Figure 1: STPS



and corr).png

Figure 2: bursts



and f).png

Figure 3: correlation

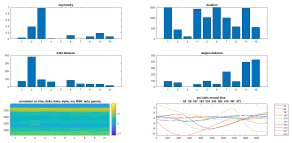


Figure 4: ictals

3 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 **??**) 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

该文档是极速PDF编辑器生成，
如果想去掉该提示,请访问并下载：
http://www.jisupdfeditor.com/

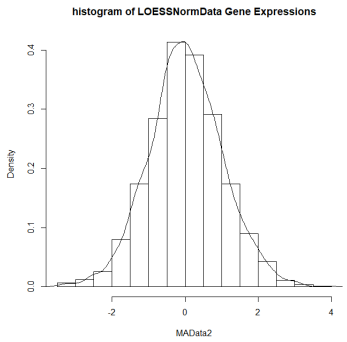
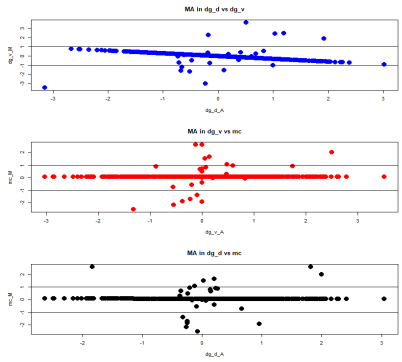


Figure 5: normalized dataset



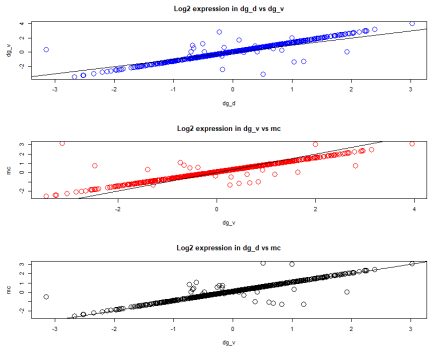
dg mc extented.png

Figure 6: MA plot

adaptability of organism in the versatility and adaptability). After the LOESS normalization, it is normalized according to QC test(5) 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.(6). Detailedly, 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 multicellular 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: 8. The cell with different markers can be seen with their

该文档是极速PDF编辑器生成，
如果想去掉该提示,请访问并下载：
http://www.jisupdfeditor.com/



expression dg mc extented.png

Figure 7: log2 after fdr correction

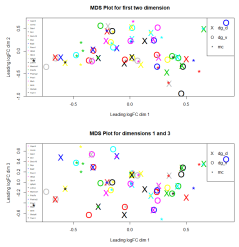


Figure 8:

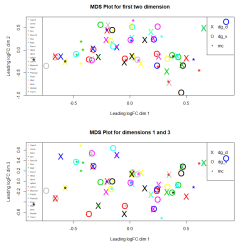


Figure 9: MDS on the first two dimensions

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 1 and cell 3, namely dg-d and mc. To be more specific, as shown in 9 marckl1 is accumulated more centrally on dimension 2 than dimension 1(along with the extended synthetic dataset 31 and 47) expressed higher on dimension 1 only in cell 2. Similarly, for dimension 1 and 3, the Prox1 and St18 are expressed evenly on both dim 1 and

该文档是极速PDF编辑器生成，
如果想去掉该提示,请访问并下载：
http://www.jisupdfeditor.com/

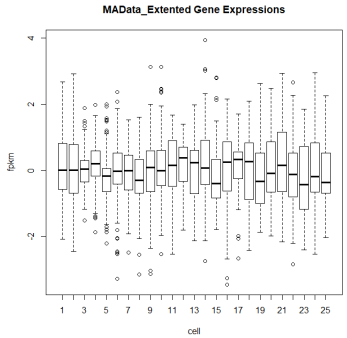
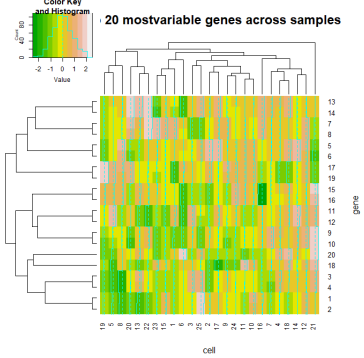


Figure 10: fpkm expression (on log level)

dim 3 while Marckl1 distributed only higher in cell 2.



heatmap of normed cells extented.png

Figure 11: top 20 expressed cell

And after studying the correlation between different genes(12), the cells and finally extended to 75 datasets for each cell and clustered(14) are classified based on the TF(13). The 8th, 9th, 20th, 21st, 22th, 23th and 25th are less correlated with 1st, 2nd, 7th, 11th, 15th, 20th and 24th individually. On average, the genes are more correlated positively rather than negatively. The extended TF is shown with the heat-map, giving rise to the density based cluster of top 20 genes across the 75 datasets (11). The highest expressed genes(negatively) are the first 4 genes in 19th, 5th, 8th and 20th dataset and the lowest for those genes are in the 6th, 3rd, 9th, 24th, 4th and 14th datasets are the first 4 genes in 19th, 5th, 8th and 20th dataset and the lowest for those genes are in the 6th, 3rd, 9th, 24th, 4th and 14th datasets while the positively highest expressed genes, 15th, 9th, 10th and 20th are all on the 21st dataset. The top 5 clusters(with lowest variant) in the dendrogram is marked in red, green yellow, blue and purple.(With dataset 1 47 72 22 31 56 6 , 39 14 64 28 3 53, 54 4 29, 27,2,52 and 30, 3, 55.) As cells can be classified by position of maximum changes, next

该文档是极速PDF编辑器生成，
如果想去掉该提示,请访问并下载：
http://www.jisupdfeditor.com/

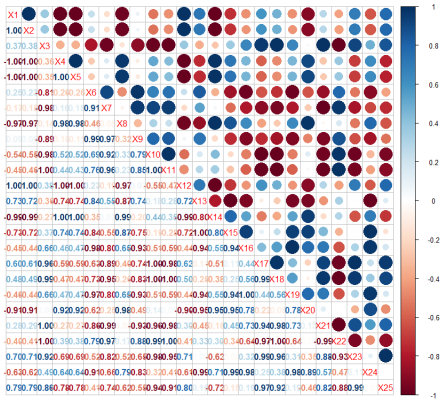


Figure 12: heatmap

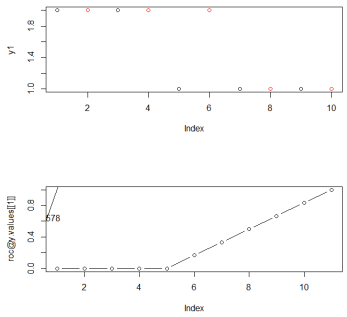


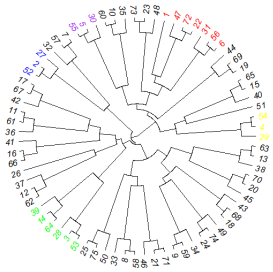
Figure 13: one classification example

study will be about the transcriptions of the cells giving out the relationships of the expression changes both temporarily and spatially.

4 t-SNE clustering with modified BIC

As the structure information is also quite useful in dna or protein alignment, Bayes Inference can also be applied on alignment combining diffusive structure information which is modeled with Hidden Markov combining RNN**?**as sequence data.X_n. The predicted dna/proteinY_n is processed based on simialrly again forward status $\alpha_{cell} = P(Y_{s_j};\theta)$, and backforward status: $\beta_{cell} = P(Y_{t+1:t|s_{j+1}};\theta)$ and the posterior is thus $\gamma = \frac{\alpha_t * \beta_t}{P(t)}$

该文档是极速PDF编辑器生成，
如果想去掉该提示,请访问并下载：
http://www.jisupdfeditor.com/



of normed genes(extended).png

Figure 14: dendrogram of the clusters

. Since our interest here is to show the advantage of the model with high dimensional RNA sequencing data, this work is not introduced here.Instead, we are going to discuss the scRNA sequencing problem with the t-SNE with modified BIC.

As the RNA sequence data are usually of high dimension, instead of MCMC, the variational method is usually utilised. Variational inference can be regarded as the optimization problem whcih thus is of lower computation consumption. The criteria is usually based on the minimisation of the Kullback-Leiber(KL) divergence , the log difference between observed and approximated posterior distribution. The method we use is modified on t-Distributed Stochastic Neighbor Embedding(t-SNE) which converts pairwise distances in high dimensional space with data points x_i, to corresponding embedding points y_i pairwise join distributions in low dimensional spaces, which respectively follows:

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

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_j is optimized through bisectional search automatically with the pre-specified perplexity $Perplex(p_j) = 2^{H(p_j)}$, where $H(P_j) = -\sigma_j p_{i|j} * log_2^{i|j}$ ao 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:

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

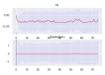
with inspiration of the Bayesian Information Criteria(BIC):

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

该文档是极速PDF编辑器生成，
如果想去掉该提示,请访问并下载：
http://www.jisupdfeditor.com/

, 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). 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_j 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.

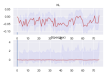
With the practical test, we apply it on the MC GC cell expression classification with t-SNE



and S(perplex) 10.png

Figure 15:

Figure 17:



and S(perplex) 173.png

Figure 16:

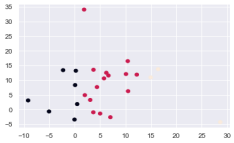
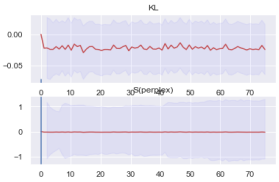
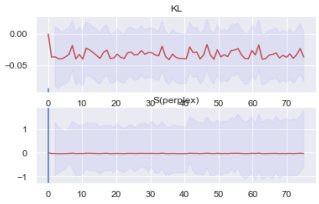


Figure 18:

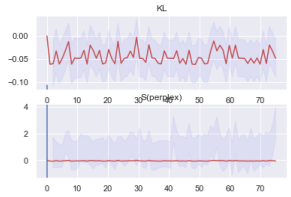


and S(perplex) 30.png



and S(perplex) 31.png

Figure 19:



and S(perplex) 32.png

该文档是极速PDF编辑器生成，
如果想去掉该提示,请访问并下载：
http://www.jisupdfeditor.com/

The top figure left is the trajectory of the gcs and mc classified under 10 groups while the top right figure gives the result classified with 5 groups and the perplex are given as 8 and 6 separately as can be seen in the bottom figure.(The optimization is based on KL minimization with fast-gradient search) The modified BIC v.s.perplexity shows the best classification is with 10 classifications with perplexity being 8. Both of the figures are shown on the first 2 dimensions of the embedding and it can be observed that since the reduction of the perplex is sharper, the classifications are also more easily separable non-linearly (closer within one classification while more alienated with inter classifications.) In fact, the standard deviation are all not too large along the whole tested perplex domain showing the good-to-fitness of the t-SNE which is as well of certain stability. As it being said, t-SNE performs generally well on high dimension classification with the modified BIC.

5 Conclusion and discussion

In summary, according to 21 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. 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 some mean on log2 level. 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.) 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.22 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.) It might be still studied on log2 level in the further exploration.

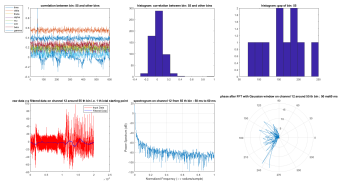


Figure 21:

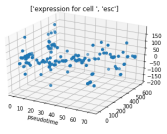
for esc.png 

Figure 22: