Classification of Synapses Using Spatial Protein Data

Jenny Chen and Micol Marchetti-Bowick

CS229 Final Project • December 11, 2009

1 MOTIVATION

Many human neurological and cognitive disorders are caused in part by aberrancies in the composition and volume densities of synapses in the brain [1]. The ability to analyze the underlying causes of such diseases would be greatly enhanced by detailed knowledge about the ratio and quantities of different types of synapses that are present in specific regions of a patient's brain. In order to make this possible, neuroscientists at Stanford and elsewhere are working towards developing methods to accurately locate and classify individual synapses within a sample of brain tissue [2]. Although many advances have been made in this area of research, synapse classification still requires biologists to manually analyze each synapse in order to evaluate its type. Here we propose a computational approach to synapse classification based on spatial protein data. Not only would the success of this method this greatly increase the efficiency of labeling synapses by type, it could also help to elucidate novel types of synapses that remain undiscovered as of yet.

2 OVERVIEW

Our goal was to use spatial information (location and density) for multiple distinct proteins within a section of brain tissue to classify "potential synapses" into one of three categories: excitatory synapses, inhibitory synapses, or non-synapses. In order to develop an effective solution to this problem, we aimed to answer the two following questions:

(1) How can we develop a model that represents a synapse accurately enough to enable us to infer its type?, and

(2) Which among the proteins from our dataset provide the most valuable information for classifying synapses and what does this imply about their biological function?

3 DATA

We obtained data from the Smith Laboratory in the Department of Molecular and Cellular Physiology at Stanford. The dataset consists of 200 labeled examples, each of which is either an excitatory synapse, an inhibitory synapse, or not a synapse. The information given for each synapse is a series of protein density readings that directly map to *(x,y,z)* locations within a 1-μm^3 cube centered at each synapse. These readings were provided for a total of 11 different proteins for every single synapse.

4 MODELS

In order to be useful for type classification, our model needed to incorporate the distinguishing characteristics of each of the three different categories of synapse.[1] One such characteristic is the relative amounts of different proteins that are present in the region around a synapse. We also hypothesized that the shape and size of a synapse, including the relative locations of the regions within the synapse where each protein is most highly concentrated, would be important indicators of its type. Based on these traits, we came up with two different models to test.

4.1 Baseline Model

The first model, which we called our "baseline" model, represents synapses of a given type as a multinomial distribution over each of the *n* proteins in the dataset. The values of the parameters of this multinomial represent the likelihood of encountering one unit of a particular type of protein in a narrow region around the synapse. If *x* is a particular unit of protein density, and *y* is the type of the synapse in question, then we have

$$P(x \mid y) \sim Multinomial(\phi_{y_1},...,\phi_{y_{n-1}})$$

where ϕ_{y_i} is the likelihood that unit *x* originates from protein *i* given that the synapse is of type *y*. Thus, a particular synapse is modeled by the relative quantities of the *n* proteins that it contains. This model only incorporates the first of the characteristics in which we are interested, but it has the advantage of simplicity.

4.2 Complex Model

The next model strives to capture the characteristic shape and size of each type of synapse. In order to accomplish this, we decided to represent each synapse by a multivariate Gaussian that models its physical location and shape in either 2-D or 3-D space (we tried both). We

[1] Here we refer to a "non-synapse" as a category of synapse even though it technically is not a synapse at all.

decided to experiment with three different variations of this idea:

1. Model each synapse as a single Gaussian.
2. Model each synapse as a composition of *n* distinct Gaussians, one for each protein.
3. Model each synapse as a composition of two distinct Gaussians, one representing the presynaptic region and the other representing the postsynaptic region.

The motivation for these three slightly different models came from thinking about the physical structure of a synapse. The most basic variant views each synapse as a single concentrated blotch of protein and models a synapse of a given type as a Gaussian over the location of each unit of protein density, so that we have

$$P(x \mid y) \sim Normal(\vec{\mu}_y, \Sigma_y)$$

where $P(x \mid y)$ is the total likelihood of all protein readings from synapse x given that the synapse is of type y (here we assume that protein readings are independent when conditioned on y). The next variant models each protein blotch from a distinct source as a separate Gaussian, as shown in Figure 1(a). The third variant is based on a slightly more sophisticated understanding of synapse structure. In reality, a synapse is not just one or several "blotches" of protein. Instead, it is composed of a presynaptic region and a postsynaptic region that are responsible for carrying out different functions. Because of this, different proteins localize to each region, which allows us to distinguish one from the other in our data. We based the third variation of our model, shown in Figure 1(b), on this inherent biological synapse structure.

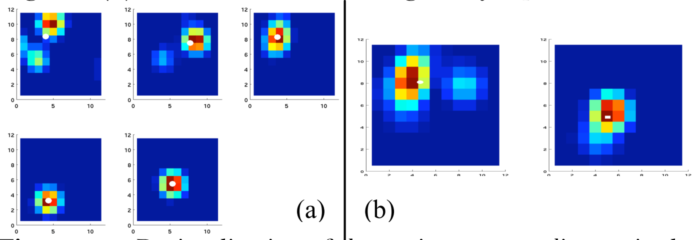

Figure 1. 2-D visualization of the region surrounding a single synapse. Red pixels indicate stronger protein readings (i.e. protein is very dense), blue pixels indicate weak protein readings. Each image also shows a white point indicating the mean location of all protein units in that image (location weighted by pixel brightness). (a) Separate protein density images for each of 5 proteins; from left to right, top to bottom, we have: *gad*, *bassoon*, *synapsin*, *gephyrin*, *PSD*. (b) Protein density images for the presynaptic region (left) and the postsynaptic region (right). The presynaptic image is made up of the cumulative densities of 3 presynaptic proteins: *gad*, *bassoon*, *synapsin*. The postsynaptic image is made up of the cumulative densities of 2 postsynaptic proteins: *gephyrin*, *PSD*. Notice that the presynaptic and postsynaptic regions are slightly offset.

5 METHODOLOGY

Multinomial logisitic regression and multinomial Gaussian discriminant analysis (GDA) were used to fit and test our various models. Because of our small data set size, we used leave-one-out cross validation (LOOCV) to estimate the generalization error of our models.

5.1 Logistic Regression & Feature Selection

We began by using logistic regression to compare our baseline and complex models. In order to test our baseline model, we calculated the normalized protein distribution of each synapse in our data set and trained our logistic regression classifier on these *n* features (one for each protein).

We next wanted to determine which aspects of our complex models contributed the most useful information to the classifier. To do this, we first applied each of the three models to every individual synapse in order to learn a set of parameters for that synapse. From these parameters, we extracted five representative features that we hypothesized would be valuable in classifying the synapses into types:

1. Variance of the locations of proteins weighted by density, averaged across all dimensions of the data (e.g. in 2-D, we take the average of var(x) and var(y)), for each of the 11 proteins (*n* features, 1 per protein). This is intended to be a measure of how diffuse each protein is within the synaptic region.
2. Variance of the mean locations of each protein (e.g. var($\mu_1,\ldots,\mu_n$) where μ_i = mean location of protein i weighted by density at each location) (1 feature).
3. Variance of the locations of presynaptic proteins and variance of the locations of postsynaptic proteins, both weighted by density and averaged across all dimensions of the data. This is analogous to feature 1 but for the pre- and post-synaptic regions (2 features).
4. Variance of the mean locations of the presynaptic proteins and variance of the mean locations of the postsynaptic proteins (2 features).
5. Difference from the average distance between the presynaptic mean and the postsynaptic mean (e.g. compared to the average, how far is this synapse's presynaptic mean from its postsynaptic mean?) (1 feature).

The first two features model the synapse as a composition of *n* proteins, while the last three features model the synapse as a composition of presynaptic and postsynaptic regions.

We ran logistic regression using each of these features in conjunction with the baseline feature (normalized protein distribution) and compared the results. We also performed each test using both 2-D features (i.e. features extracted along a fixed z-plane) and 3-D features to compare the information gained in three dimensions.

We found that many of the above features contributed little or nothing to the accuracy of the classifier. The feature that helped the most was the measure of how diffuse each protein is within the synaptic region (feature 1). In addition, we discovered that models that used 3-D features consistently matched or outperformed models that were based on 2-D data. Finally, we found that using smaller regions to extract features significantly improved the accuracy of our model (Figure 3).

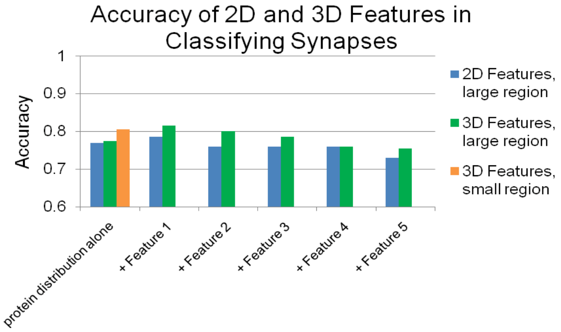

Figure 3. Accuracy of multinomial logistic regression classifier using each of the six 2-D features and 3-D features in conjunction with protein distribution alone. Feature 1, the measure of how diffuse each protein is, contributed the most accuracy out of all the features. Additionally, we see from column 1 that 3D features extracted from a "small region" (a 7x7x7 pixel area) provided more accuracy than the "large region" (11x11x11 pixels).

5.2 Gaussian Discriminant Analysis (GDA)

To further explore our complex models, we implemented a GDA algorithm to fit parameters to our data for each of the three models:

Model 1: We fit a single multivariate Gaussian to each of the three types of synapses.

Model 2: We modeled a separate Gaussian over each protein to characterize each synapse type by a set of *n* Gaussians.

Model 3: Finally, we learnt a presynaptic Gaussian and a postsynaptic Gaussian for each synapse type.

After learning each model, we classified new synapses by calculating $P(x|y)$, the probability that synapse x was generated by the Gaussian model(s) with parameters $\mu_{y\_1},\ldots,\mu_{y\_d}$ and $\Sigma_{y\_1},\ldots,\Sigma_{y\_d}$ (where $d = 1$ for model 1, $d = n$ for model 2, $d = 2$ for model 3) and $P(y)$, the probability of encountering a synapse of type y. We calculated $P(x|y)P(y)$ for every y and selected the type category with highest probability.

We used three-dimensional data for all three implementations and restricted our analysis to using the five proteins *gad*, *bassoon*, *synapsin*, *gephyrin*, and *PSD* (the first three of which are presynaptic proteins, and the last two of which are postsynaptic proteins). We obtained the best results using model 2, which maintains a separate representation for the regions of density of each protein, rather than massing them together (Figure 4). However, neither of the three achieved a particularly high accuracy.

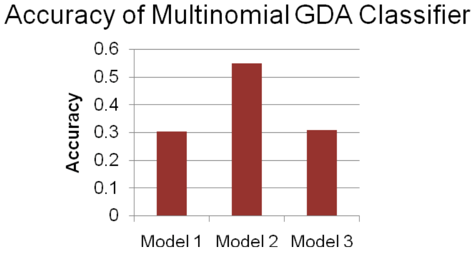

Figure 4. Accuracy of multinomial GDA classifier using each of the three complex models. Model 2, which represents each protein separately, was the most successful.

6 RESULTS

Our most successful classification was realized using logistic regression operating on the baseline protein distribution along with feature set 1 (measures of how diffuse each protein is) in three dimensions. Our model yielded 84% accuracy with an ROC AUC of 0.79 for non-synapse/synapse classification and 0.96 for excitatory /inhibitory synapse classification. In comparison, logistic regression operating only on the baseline model of 2-D protein distribution yields 77% accuracy with an ROC AUC of 0.71 for non-synapse/synapse classification and 0.93 for excitatory/inhibitory synapse classification (Figure 4).

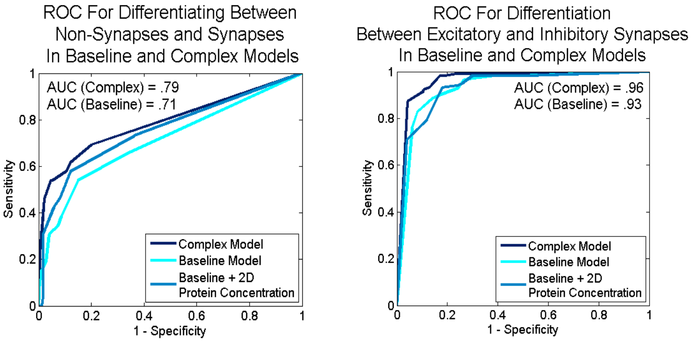

Figure 4. Final ROC curves comparing baseline and complex models in differentiating between non-synapses and synapses (left) and differentiating between excitatory and inhibitory synapses (right). Here, "Protein Concentration" refers to Feature 1 in 2D.

7 BIOLOGICAL RELEVANCE

Though we began with feature information for 11 proteins, this was far too much information to train on with only 200 data points. Using forward search, we determined a subset of four proteins that yielded maximal accuracy: *bassoon*, *synapsin*, *gad*, and *PSD*. We trained and tested our model using each protein alone to infer the relative importance of each protein (Figure 5). *synapsin* was found to be the best at distinguishing between non-synapses and synapses and *gad* was the most useful for distinguishing between excitatory and inhibitory synapses. These results agree with the currently known functions of these proteins. *synapsin* is thought to be present in every synapse, which explains its strong performance in distinguishing between synapses and non-synapses, and *gad* is known to be a good marker of inhibitory synapses [3,4].

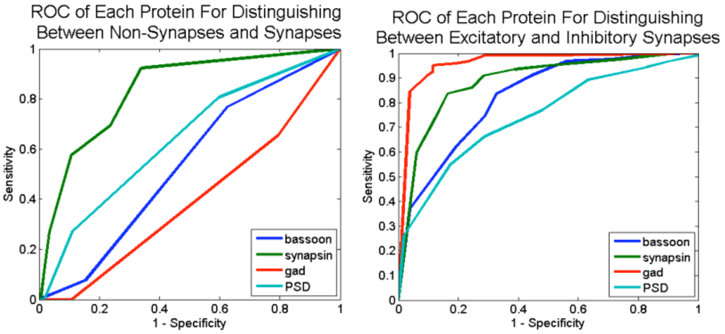

Figure 5. ROC of each protein alone in ability to distinguish between non-synapses and synapses (left) and between excitatory and inhibitory synapses (right).

8 DISCUSSION

Our results show that the protein distribution at a synapse is by far the best predictive characteristic of the data that we have for synapse type. This result is unsurprising because it is based in biological fact: proteins have very specific biologic functions and localize to specific synapse types. Here, we also explored other features of the data that would enrich our synapse model and improve our classification accuracy. We attempted to represent the shape and size of synapses in a variety of ways. We included certain shape/size features as input to our logistic regression classifier, and we also implemented a multiclass GDA algorithm that modeled each class (i.e. each synapse type) as a different Gaussian or set of Gaussians. Although not all of these approaches were successful, our results clearly demonstrate that the shape and size of a synapse are relevant to its type. We have two results that lead us to this conclusion. First, we saw that when we added a measure of how "diffuse" each protein is at the synapse as an input feature to our logistic regression classifier, our accuracy improved. Secondly, our GDA classifier was most successful when using a model that represented the density units from each protein as a separate Gaussian distribution.

In general, our GDA classifier was less successful than our logistic regression classifier. We believe that the principal reason for the discrepancy between the two is that our GDA model did not explicitly include a representation of the relative amounts of each protein present within each type of synapse, which has proven to be a very powerful feature. However, it is interesting to note that GDA using complex model 2 was still quite successful even without this feature.

Although we tested many features in our logistic regression classifier, only a few actually contributed to improving our accuracy. One model in particular that seemed to be of little worth is the pre-/post- synaptic region model. The features extracted from this model did not help with logistic regression, nor did it produce good results when used as a basis for GDA. One possible reason for this is that the pre-/post- synaptic regions are not well represented within the data. The resolution may actually be too low to capture a clear relationship between the two regions.

There were many other complicating factors that made it difficult to extract useful features and construct an accurate synapse model. For example, it is possible that some of the labeled regions we were working with actually contained multiple synapses of different classes. These mixed signals certainly contributed to our inability to achieve above a maximum of 84% accuracy.

Finally, we were constricted by the fact that we had very few training examples to work with; we only had a total of 200 labeled synapses. Furthermore, within this set, less than 1/5 of our examples were non-synapses, which could explain why our ability to distinguish synapses from non-synapses is much poorer than our ability to distinguish excitatory synapses from inhibitory ones.

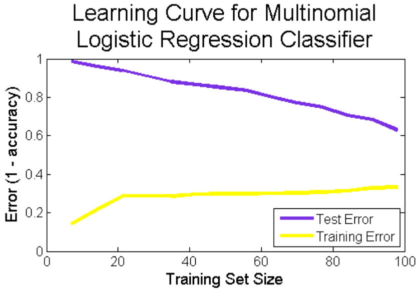

Figure 6. Learning Curve for multinomial logistic regression classifier trained on our complex model.

Another consequence of our small dataset was that our classifier has fairly high variance, indicated by the fact that the test error is much higher than the training error (see learning curve plotted in Figure 6). However, we believe that this could be easily fixed with a larger training set size.

It is worth noting that labeling of synapses by hand is a difficult problem in itself and has a fairly high error rate. Therefore, our 84% accuracy rate may or may not reflect biological truth. Ideally, a computational approach such as the one proposed here could be used in conjunction to elucidate biology. For example, synapses that get labeled as non-synapses by a computational method are ideal candidates to be reexamined by a neuroscientist to determine whether they were incorrectly labeled or perhaps belong to a novel type of synapse that does not fall into either the excitatory or inhibitory category.

9 FUTURE DIRECTIONS

The inability of any of our classifiers to surpass an accuracy level of 84% may be due in large part to the noise that exists in our data. None of the models we propose account this noise, even though it almost certainly impacts the features we use.

One important source of error is the inaccuracy which accrues from the wetlab techniques that are used to generate our data. The spatial protein data we use is collected using a technique called immunostaining, whereby fluorescence markers are bound to proteins present in the tissue. However, these markers are not guaranteed to bind evenly across all proteins, which can lead to a misrepresentation of the true underlying protein densities. For example, it is possible that *synapsin* was simply not detected at a particular location, which might lead to that region to be inaccurately classified as a non-synapse even though all other features indicate otherwise.

A possible extension to our project would be to modify and extend our model so that it takes this noise into account and is able to adjust for the possibility that the data is not 100% accurate.

Another source of complexity is the potential for synapses to overlap within the brain. Currently, the model we use would inherently interpret two overlapping synapses as a single entity, which again leads to a misrepresentation of the data. A model and inference strategy that could account for this situation as well would be all the more powerful.

Another exciting direction in which to take this project in the future would be an attempt to apply an *unsupervised* learning algorithm to cluster synapses into different types based on the features that we've found to be most salient. This approach could move beyond distinguishing between the three basic categories that we've discussed (excitatory, inhibitory, non-synapse) and potentially may reveal new classes of synapses that might be biologically significant.

REFERENCES

1. Fiala, J.C., Spacek, J., and Harris, K.M. (2002). Dendritic Spine Pathology: Cause or Consequence of Neruological Disorders? *Brain Research Reviews* 39, 29-54.
2. Micheva, K. and Smith, S.J. (2007). Array Tomography: A New Tool for the Molecular Architecture and Ultrastructure of Neural Circuits. *Neuron.* 55, 25-36.
3. De Camilli, P., Cameron, R., and Greengard, P. (1983). Synapsin I (protein I), a nerve terminal-specific phosphoprotein. I. Its general distribution in synapses of the central and peripheral nervous system demonstrated by immunofluorescence in frozen and plastic sections. *J. Cell Biol.* 96, 1337–54.
4. Soghomonian, J-J., and Martin J.L. (1998). Two isoforms of glutamate decarboxylase: why? *Trends in Pharmocol Sci.* 19(12), 500-5.