A Bayesian Kernel for the Prediction of Neuron Properties from Binary Gene Profiles

François Fleuret & Wulfram Gerstner
Laboratory of Computational Neuroscience
EPFL, Station 15
CH-1015 Lausanne, Switzerland
francois.fleuret@epfl.ch

Abstract

Predicting cellular properties from molecular or genetic data is a challenge for bioinformatics and machine learning. In brain slices of neuronal tissue, it has become possible to both measure electro-physiological properties of a given neuron and to extract a sample of its cytoplasm so that expressed genes can be amplified. Thus, the presence or absence of genes related to ion channels in the neuronal cell membrane can be correlated with neuronal behavior encoded as a set of electro-physiological parameters. A typical gene amplification process is asymmetric in the sense that false positives are very rare, whereas false negatives (genes expressed but not amplified) are rather common. An analysis of a probabilistic model of that process yields a similarity measure between two strings of amplified genes that takes the asymmetry of the amplification process into account. This similarity measure can be put under the form of a conformal-transformed kernel. We provide experiments with support-vector machines on artificial and neuronal data.

1. Introduction

The purpose of this paper is to study the prediction of the electro-physiological properties of cortical neurons from the result of the amplification of a family of genes related to ion channels. We use a non-parametric regression method known as a Support Vector Machine[7, 2] which has demonstrated great performance on a large class of problems, including inference of biological data. This method requires to chose a similarity measure on the input data space which are in our case binary strings standing for the result of the gene amplification. Each binary digit represent the presence or absence of one of the gene in the amplification.

We introduce a model of the errors appearing in the gene amplification technique and propose to use the conditional probability for the neurons to have the same expressed genes, given the amplified genes, as a similarity measure between the amplification strings. We show analytically that such a similarity measure can be put under the form of a conformal-transformed kernel, which has already been studied in the context of Support Vector Machines[5, 1].

Experiments on both real and synthetic data demonstrate that the optimal kernel parameter values are consistent with the high false-negative and low false-positive error rates of the gene amplification method. However, the improvement in prediction performance remains marginal, and the optimal results are obtained with the symmetric form of the conformal-transformed kernel.

In §2 we describe how the gene amplification is done, and our probabilistic model of the amplification errors. We present in §3 a similarity measure based on a computation of the conditional probability for the strings of expressed genes to be identical, given the two strings of amplified genes. Finally, in §4 and §5 we give and discuss experimental results.

2. Probabilistic model

We propose a model of the distribution of amplified genes, given the expressed ones: for every gene expressed, there is a high probability for its amplification to fail, while on the contrary there is a close to zero probability for non-expressed genes to be amplified by mistake.

Intuitively, this leads to an interesting asymmetry: while the detections of the same gene in two different neurons ensure that they really share this gene, the absence in both is less informative and does not prove that they share the absence of the gene, as the gene may be actually expressed in one of the neuron and was missed during amplification.

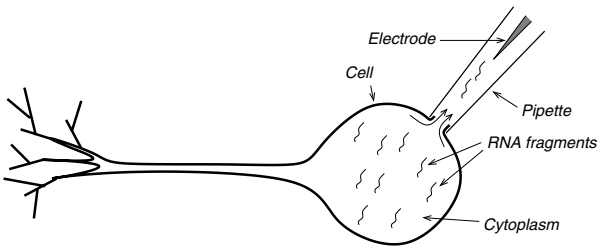

Figure 1. Patch-clamp procedure: suction breaks the cell membrane and makes the cytoplasm enter the pipette and get in contact with the electrode. RNA fragments from the cell also get into the pipette and can be amplified later on by RT-PCR.

2.1. Measure of electro-physiological parameters

The measure of electro-physiological properties of the cell is done using whole-cell patch clamp [3, 4] (see figure 1). This procedure consists of putting a hollow tube in contact with the cell and breaking the exterior membrane by suction, so that a part of the cytoplasm gets into the tube in contact with an electrode.

That electrode is then used to both inject current in the cell and measure the cell's response. The signal is processed under digital form after sampling in time and amplitude. Details about the experimental procedures are given in [6].

2.2. Gene amplification

A small volume of cell fluid is extracted from the neuron, and subject to the Reverse Transcription – Polymerase Chain Reaction (RT-PCR) which is a molecular biological method for amplifying DNA from RNA strands. The RNA is transcribed into DNA using the reverse transcriptase enzyme, and the DNA is then exponentially amplified through the PCR (Polymerase Chain Reaction) process.

The product of the PCR is then injected into a gel and subjected to an electric field. Negatively charged DNA molecules migrate in the gel to a location depending on their lengths. By comparing those locations to that of known fragments, they can be identified.

The amplification of the RNA related to one gene can fail if there are no such RNA fragment in the sample of cytoplasm used for the RT-PCR. Inversely, if alien RNA contaminates the sample of fluid, there can be false positive amplifications. However, it can be assumed that the experimental procedures prevent such contamination with high confidence.

2.3. Analytical model

In the following we denote by X the random variables on $\{0,1\}^N$ standing for the string of amplified genes (measurement), and Z the string of expressed genes (hidden truth). The value 1 stands for "expressed" or "amplified" while 0 stands for "non expressed" or "non amplified". The only information we have access to is the value of X, and we have to infer some property of Z from the stochastic relation between X and Z (see figure 2).

If $l \in \{1,\ldots,N\}$, and s is a string of N elements (or a random variable on the string space), we denote in the rest of the paper by $s^{(l)}$ the value of the l-th digit of the string s (which can be seen as the expression or the amplification status of the l-th gene).

The amplification process makes false positive (amplification of non-expressed genes, for instance by contamination by alien DNA) very unlikely, while the probability ϵ of false negatives (non-amplification of expressed genes, for instance because no RNA was caught) is high. If we consider a null false-positive error rate, this leads to the following model of $P(X^{(l)} \mid Z^{(l)})$, where ϵ is the false-negative probability:

$$\begin{aligned} P(X^{(l)}=0 \mid Z^{(l)}=0) &= 1 \\ P(X^{(l)}=0 \mid Z^{(l)}=1) &= \epsilon \\ P(X^{(l)}=1 \mid Z^{(l)}=0) &= 0 \\ P(X^{(l)}=1 \mid Z^{(l)}=1) &= 1-\epsilon \end{aligned}$$

This expresses the fact that there is zero probability for the gene to be amplified if it is not actually expressed in the cell, while it has a probability ϵ not to be detected, even if it is in the cell.

To complete our model, we also make the assumption that the amplification errors are independent, which can be formulated as the conditional independence of the $X^{(l)}$, given Z. This assumption is legitimate, considering that the false negative errors are mostly due to the absence of RNA fragments in the fluid sample taken from the cell. To illustrate that point, consider a bag containing a very large number of white balls (the cytoplasm) and a few groups of balls of several other colors (RNA fragment related to certain genes). If the number of sampled balls is large relatively to the size of the colored ball population, picking balls of one color does not influence the probability to pick balls of another color, thus making the events of picking balls of certain colors independent.

3. A gene-based kernel

One of the generic tools most used in statistical learning today are the so-called "kernel methods" [5]. These tech-

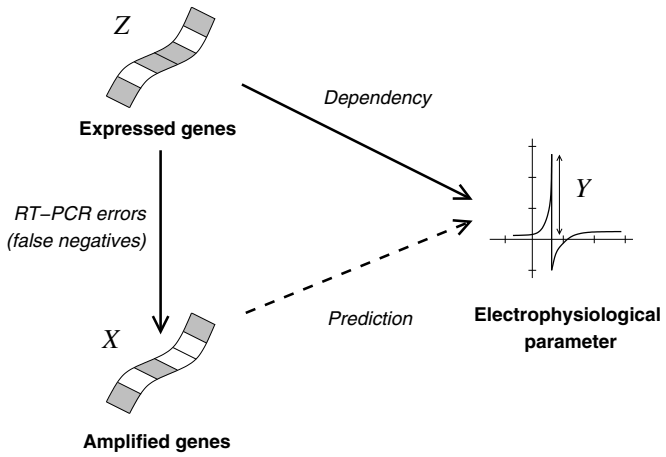

Figure 2. In our model, the electrophysiological parameter Y we want to predict is a function of the string of expressed genes Z, which is indirectly known through a stochastic and noisy string of amplified genes X. Even if the parameter to be predicted were a deterministic function of the expressed genes, the prediction could not be exact since information is lost during the amplification.

niques generalize linear methods such as regression or PCA by mapping the data into a space of large dimension beforehand.

The resulting generalized dot product is often referred to as a *Kernel*, and can be seen as a similarity measure. We describe in §3.1 the classical Support-Vector Machine technique, which we use for our experiment.

Based on the probabilistic formulation of §2.3, we derive a kernel that can play the role of a similarity measure between two strings of amplified genes.

3.1. Regression with SVMs

While Support Vector Machines were originally developed for classification (i.e. predict a discrete value from known data), they have demonstrated great performances when used for regression. In the linear case, the predictor $f(X)$ has the form of a linear function of the input vector $f(x) = \langle x,\, \omega \rangle$. The training process selects an ω with good generalization properties by minimizing

$$E(\omega) = \sum_i \mid y_i - \langle x_i,\, \omega \rangle \mid_\epsilon \, + \, ||\omega||^2$$

where $|e|_\epsilon$ is $e - \epsilon$ for $e > \epsilon$ and 0 else. This cost function forces the predicted value to be at less than ϵ than the training ones, and it can be minimized optimally.

This rule can be improved by combining it with a mapping Φ in a space of higher dimension. Formally, the mapping Φ does not have to be explicit, since it only appears in dot products of the form $\langle \Phi(x),\, \Phi(x') \rangle$. Thus, any methods relying on a linear rule can be generalized into a "kernelized version" as soon as one is provided with the expression of $k(x, x') = \langle \Phi(x),\, \Phi(x') \rangle$, even if neither Φ nor even the high-dimension space have been made explicit.

However, this kernel k keeps the role of a similarity measure it has in the basic linear case.

3.2. Conditional probability as a kernel

Given two strings x_1 and x_2 of amplification results, we propose to quantify the similarity between the strings as the probability for the expressed genes to be the same in both neurons, given that the two strings of amplified genes are respectively x_1 and x_2:

$$k(x_1,\, x_2) = P(Z_1 = Z_2 \mid X_1 = x_1,\, X_2 = x_2)$$

This value can be evaluated with a simple Bayesian rule. We know that X_1 and X_2 are independent, and that Z_1 and Z_2 are independent too. Also, according to our model, the $X_k^{(l)}$ are conditionally independent given Z_k. We can prove that

$$k(x_1,\, x_2) = \prod_{l=1}^{N} \kappa_l(x_1^{(l)}, x_2^{(l)})$$

with:

$$\kappa_l(a, b) = \sum_{c \in \{0,1\}} P(Z_1^{(l)} = c \mid X_1^{(l)} = a)\, P(Z_2^{(l)} = c \mid X_2^{(l)} = b)$$

Note that κ_l can be interpreted as a similarity measure between neurons based on the presence or absence of the k-th gene alone. It will take into account the high false negative rate and the absence of false positive. We define the following quantities:

$$\begin{aligned} f_n &= P(X^{(l)} = 0 \mid Z^{(l)} = 1) \\ f_p &= P(X^{(l)} = 1 \mid Z^{(l)} = 0) \\ \alpha &= P(X^{(l)} = 1) \\ \beta &= P(Z^{(l)} = 1) = \frac{\alpha - f_p}{1 - (f_p + f_n)} \end{aligned}$$

From which we compute κ_l:

$$\begin{aligned} \kappa_l(0,0) &= \frac{1}{(1-\alpha)^2} \left((1-\beta)^2 (1-f_p)^2 + \beta^2 {f_n}^2 \right) \\ \kappa_l(1,0) &= \kappa_l(0,1) \\ &= \frac{1}{\alpha(1-\alpha)} \left((1-\beta)^2 f_p (1-f_p) + \beta^2 f_n (1-f_n) \right) \end{aligned}$$

$$\kappa_l(1,1) \quad = \quad \frac{1}{\alpha^2}\left((1-\beta)^2 {f_p}^2 + \beta^2(1-f_n)^2\right)$$

3.3. Conditional probability as a conformal-transformed kernel

Since this kernel is defined on binary strings, we can prove that for adequate values of δ and γ

$$k(x_1,\, x_2) \;\propto\; \exp\left(\,\delta||x_1||^2 + \delta||x_2||^2 + \gamma\,||x_1 - x_2||^2\,\right),$$

which is a conformal-transformed kernel [5, 1] if $\gamma < 0$. The coefficient δ is positive if the presence of an amplified gene is more informative than its absence ($\kappa(1,1) > \kappa(0,0)$), and negative otherwise. Thus, if there are more false-negatives than false-positives, samples with a lot of amplified genes (i.e. large $||x_1||^2$) are more significant and more weighted.

4. Experiments

We propose to validate the results presented above by training and testing SVMs on both synthetic and real data. We will look at the performance of a standard linear predictor (i.e. classical linear regression), regularized linear predictor and the custom quasi-conformal kernel of the form presented in §3.3.

All the experiments have been done with softwares written in C++ on GNU/Linux computers. We have used free software tools (editor, compiler, debugger, word-processors, etc.), mainly from the Free Software Foundation[1]. We have also used the Libsvm[2] from Chih-Chung Chang and Chih-Jen Lin.

4.1. Synthetic data

To test the efficiency of this new kernel, we have generated synthetic data according to our model presented in §2 consisting of simulating the amplification of 9 genes in 100 neurons and a virtual EP function. Each gene expression has marginal probability 0.5. The 9 genes are divided into two sub-groups (one of 5 genes, the other of 4) and the EP functions depends on the number of genes expressed in each of those groups. If more than half of the genes of the first group are expressed, than the EP is equal to the number of genes expressed in the second group, if less than half of the genes of the first group are expressed, than the EP is equal to 4 minus the number of genes expressed in the second group. Such an EP is simple, yet highly non-linear.

[1] `http://www.fsf.org`

[2] `http://www.csie.ntu.edu.tw/~cjlin/libsvm`

The amplification is simulated by flipping at random the simulated expressed genes. In the first experiments, the flipping is done symmetrically with probability 0.1 for both false positive (i.e. genes non expressed which are amplified) and false negatives (genes actually expressed but not amplified). In the second experiment, false positives occur with probability 0.01 while false negatives occur with probability 0.2, which is more similar to real experiments.

Optimal Bayesian predictor: Because we know the true distribution of the data, we can compute the conditional expectation, given the amplified genes. This value is optimal for the quadratic error.

The value computed by this predictor for a given string of amplified genes $x^\star$ can be understood intuitively as the following: let's imagine that we generate a very large – virtually infinite – number of synthetic experiments. For each of them, we generate a string z, from which we compute the EP, and a second string x of amplified genes, obtained by flipping off certain expressed genes. If the x thus obtained is equal to $x^\star$, we note the value of the EP. The average of all those collected EP values is the conditional expectation, and our prediction. In practice, the computation is not done that way. Instead, we analytically compute the conditional distribution on the values of the EP, and from it, we compute the expected value.

Note that this Bayesian predictor does not use any training set: it directly estimates the averaged value to predict, given the amplified genes.

Estimation of δ: As introduced in §3.3, the δ parameter controls the asymmetry of the kernel: a positive δ will lead to more weight on examples with a large number of amplified genes, while a negative δ will put more weight on examples with a small number of amplified genes.

The optimal δ is estimated by training the SVM on a training set for different values of δ and keeping the delta leading to the optimal error rate on a validation set. Those sets are generated according to the model described above. This is repeated 100 times, each time with a training and validation sets of size 100.

Histograms of the optimal δ in the symmetric and asymmetric situation are shown on figure 3. The results are consistent with the model: the assymetry in the amplification errors leads to an assymetry in the optimal δ distribution.

Prediction performance: Prediction performance is estimated by computing the correlation between the true and the predicted value, estimated on 100 samples. We can compare two predictors by repeating the experiment 100 times, each one consisting of several prediction from which we can compute a correlation. Correlations obtained with different methods can be plotted against each others. The position of

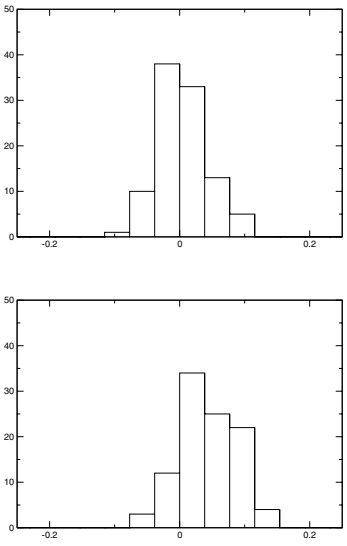

Figure 3. Top: distribution of optimal δ on synthetic data in the symmetric case $FN = 0.1$ and $FP = 0.1$, 49 out of 100 are negative. Bottom: distribution of optimal δ on synthetic data in the asymmetric case $FN = 0.2$ and $FP = 0.01$. Only 15 out of 100 are negative.

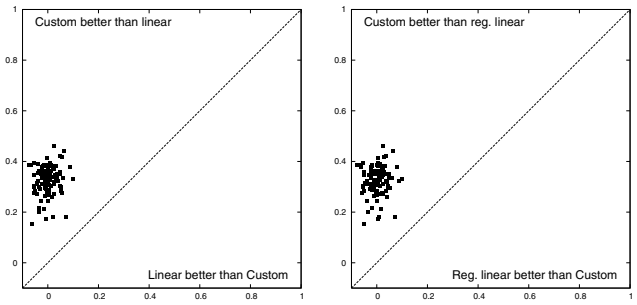

Figure 4. Custom conformal-transformed vs. linear and regularized linear on synthetic data with $FN = 0.2$ and $FP = 0.01$

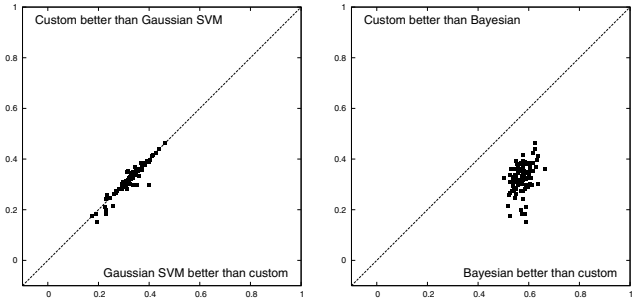

Figure 5. Custom conformal-transformed vs. Gaussian and Bayesian on synthetic data with $FN = 0.2$ and $FP = 0.01$

the plotted points with respect to the diagonal gives a good indication of the relative performances of the two methods. Results are shown on figure 4 and 5.

As expected, SVM performs better than linear or regularized linear regression, and worst than the optimal Bayesian predictor. The surprising result is the similar performance of the Gaussian and our custom conformal-transformed kernel.

4.2. Real data

Those real data consist of 183 neurons. For each neurons, we are provided with a binary vector $(x_1, \ldots, x_{29}) \in \{0, 1\}^{29}$ representing the amplification result of 29 genes, and a real-valued vector $(y_1, \ldots, y_{61}) \in R^{61}$ of 61 electrophysiological (EP) measures, as described in §2.1, §2.2. See [6] for details about this data set and the extraction process.

Estimation of δ: For each one of the 61 EP, the first set of experiments consist of estimating the optimal δ through several rounds of cross-validation to check the consistency with the analytical model we propose. Due to the high false negative error rate, the optimal δ should be positive, leading to a greater influence of training samples with more amplified genes.

Figure 6 shows how the 61 optimal δ are distributed. A very large majority are positive, and less than 17% are strictly negative.

Prediction performance: A straight-forward prediction scheme is the standard linear regression. Despite its simplicity, it behaves well in the context of noisy data and small training sets such as the one we have to deal with here.

Prediction with SVM outperforms both linear regression and regularized linear regression, see figure 7. However, the asymmetric kernel does not provide a significant improvement in prediction: as seen on figure 8, the performance are only marginally better with an asymmetric kernel and an optimized δ than with a Gaussian kernel. The most likely reason for such an absence of performance is due to overfitting of the model for small data sets, and good performance of the Gaussian kernel for large dataset. In both cases, the asymmetry does not bring additional prediction capabilities.

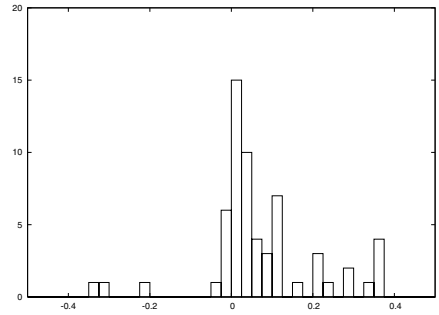

Figure 6. Of the 61 optimal δ estimated on the real data 16 are null and 35 are strictly positive. Only 10 are negative. This is consistent with a false negative error rate higher than a false positive one.

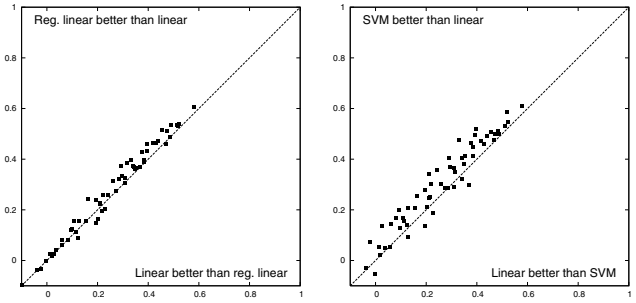

Figure 7. Correlation with the regularized linear regression vs. correlation with linear regression (left) and correlation with Gaussian SVM vs. correlation with linear regression (right).

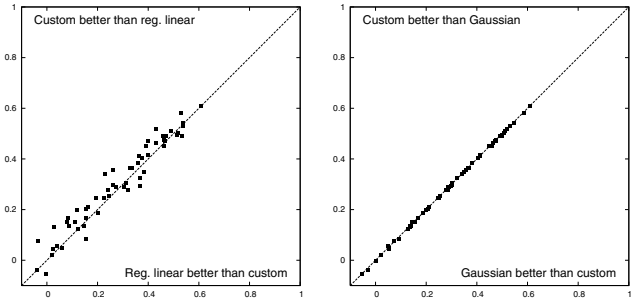

Figure 8. Correlation on the test set with the custom conformal-transformed SVM vs. correlation with regularized linear regression (left) and correlation on the test set with asymmetric kernel SVM vs. correlation with standard Gaussian SVM (right).

5. Conclusion

We proposed in this article an analytical model of the RT-PCR amplification errors, and derived from that model that the use of a pseudo-conformal kernel is sound to predict phenotypical parameter values.

The experiments demonstrate the consistency between our model and the hypothesis of false negatives in the RT-PCR. When we optimize the parameter of that kernel, the obtained values are consistent with a high false-negative error rate and may be a meaningful procedure to detect it. Experiments show that the quality of the prediction of the phenotypical values does not improve when the asymmetrical component of the kernel is optimized.

In this study, we have ignored a second source of false negative, which is the exhaustion of the polymerase during the amplification process. A few RNA strands can "take over" early in the RT-PCR process, and due to the exponential reaction can let other strands non-amplified because of a lack of polymerase later in the process. Such an effect creates strong statistical dependencies between individual gene amplifications and would lead to a more computationally expensive similarity measure.

References

[1] S. Amari and S. Wu. Improving support vector machine classifiers by modifying kernel functions. *Neural Networks*, 12:783 – 789, 2000.

[2] N. Christiani and J. Shawe-Taylor. *An Introduction to Support Vector Machines and other kernel-based learning methods*. Cambridge University Press, 2000.

[3] O. P. Hamill, A. Marty, E. Neher, B. Sakmann, and F. J. Sigworth. Improved patchclamp techniques for high-resolution current recording from cells and cell-free membrane patches. 391:85–100, 1981.

[4] B. Sakmann and E. Neher. *Single Channel Recording*. Plenum, 1983.

[5] A. J. Smola, B. Scholkopf, and C. J. C. Burges. *Advances in Kernel Methods*. MIT Press, 1998.

[6] M. Toledo-Rodriguez, B. Blumenfeld, C. Wu, J. Luo, B. Attali, P. Goodman, and H. Markram. Correlation maps allow neuronal electrical properties to be predicted from single-cell gene expression profiles in rat neocortex. *Cerebral Cortex*, 14(12):1310–1327, 2004.

[7] N. Vapnik, V. *The Nature of Statistical Learning Theory*. Springer-Verlag, 1998.