35th Annual International Conference of the IEEE EMBS
Osaka, Japan, 3 - 7 July, 2013

Experimental Comparison of Classification Methods for Key Kinase Identification for Neurite Elongation

Yuji Yoshida[1] , Kei Majima[2] , Tatsuya Yamada[2] , Yuki Maruno[2] , Yuichi Sakumura[1,3] and Kazushi Ikeda[2]

Abstract— **Kinases in a developing neuron play important roles in elongating a neurite with their complex interactions. To elucidate the effect of each kinase on neurite elongation and regeneration from a small set of experiments, we applied machine learning methods to synthetic datasets based on a biologically feasible model. The result showed the ridged partial least squares (RPLS) algorithm performed better than other standard algorithms such as naive Bayes classifier, support vector machines and random forest classification. This suggests the effectiveness of dimension reduction done in RPLS.**

I. INTRODUCTION

Any biological system can show various functions by complex interactions among biological molecules. A developing neuron also forms protruding objects called neurites by the network of various molecules: kinases, phosphatases, cytoskeletal molecules and chemical elements such as calcium ions. If the mechanisms of the network for neurite elongation and regeneration would be elucidated, new therapies targeting it might be possible for some clinical conditions associated with a loss of axon and/or dendrite connectivity such as spinal cord injury, traumatic brain injury, and apoplexy.

Of all the molecules related to regulating neurite elongation, kinases play the most important roles [1], [2], [3]. In fact, some kinases enhance neurite elongation and others inhibit it. If the effectiveness of each kinase is elucidated, the neurite elongation can be controlled for medical purposes.

There are two biological approaches to solve this problem. The one is gene manipulation, where a single kinase is inhibited one by one from hundreds of kinases to find effective combinations of kinases [4], [5]. This is costly and time-consuming. The other is pharmacological application, where a chemical drug is used for controlling neurite length [8]. Since a chemical drug inhibits a set of kinases simultaneously, this is more efficient to find effective combinations. However, it is still unclear which set of kinases should be inhibited by how much levels.

The problem of finding effective combinations of kinases from the experimental observations of chemical drug applications is mathematically ill-posed. For example, Broad Connectivity Map compound database [6] provides about $n = 1300$ profiles of drug application, which is much smaller than the number of genes ($p > 9$ million) [7] Hence, the effectiveness of each kinase can not be determined uniquely.

To ease the problem above, this paper formulates it as a classification problem. That is, the task is to determine whether a drug has the elongation effect or the shrinkage from a given set of drug-effect pairs. Nevertheless, the problem is still difficult and naive method such as support vector machines (SVM) does not work well as shown later due to high-level noise and high dimensionality of the kinase space.

One approach to avoid such degradation is dimension reduction. For similar structure data, the partial least squares method (PLS) has been applied [15], [17], [18], [19]. PLS is similar to principal component regression (PCR) [20], [21] but it performs a regression in a score space instead of the raw data space. By virtue of the property, PLS performs well even when the dataset has a strong correlation in variables that are kinases in our case. Actually many biological molecules are similar in their functions. Hence, our algorithm is based on PLS.

In the preceding study [22], a method combining PLS and ridge penalized logistic regression (RPLS) was applied for data-mining of gene microarray data in molecular biology. The microarray data also has small samples and large number of genes. In this study, we applied RPLS to our synthetic datasets including kinases and neurite lengths in order to know how well it works. The synthetic datasets were made using a mathematical model of the biochemical reactions since natural data have no ground truth. To examine the efficiency of RPLS, we compared the result by PLS with those by other statistical methods including naive bayes classifier (NBC), SVM, and random forest classifier (RFC). The result showed the superiority of RPLS for the kinase data, suggesting that the dataset contains the correlation and/or some structures among kinases.

II. METHOD

A. Dataset

The interactions between biological molecules are biochemical reactions. A primary protein-protein interactions can be described as the state transition between monomer and binding states. Although the detailed processes for the enzymatic reaction associated with kinase and phosphatase are so complicated that three types of state transitions are necessary, linearization of the reaction gives insight to its fundamental properties.

The simplest biochemical reaction equations for enzymatic reaction is a balance between increase and decrease rates,

[1] Graduate School of Biological Science, Nara Institute of Science and Technology

[2] Graduate School of Information Science, Nara Institute of Science and Technology

[3] Graduate School of Information Science and Technology, Aichi Prefectural University

978-1-4577-0216-7/13/$26.00 ©2013 IEEE

that is,

$$\frac{da_i}{dt} = k_i^f(1-a_i) - k^b a_i, \qquad (1)$$

where a_i is the normalized concentration of the ith kinase in the phosphorylated state, and thus $1-a_i$ is that the dephosphorylated state (Fig. 1A). The parameters, k_i^f and k^b, are the forward (phosphorylation) and backward (dephosphorylation) reaction rates, respectively.

Here we introduced two assumptions on kinases. The one is that each kinase functions as an enzyme (active state) when it is phosphorylated or dephosphorylated. The other is that the phosphorylation level of each kinase is regulated by other active kinases while dephosphorylation is done by phosphatases with the constant rate. The assmptions means that only the forward reactions are modeled as functions of the other kinases as

$$k_i^f = \sum_{j \in \text{act}, j \neq i} k_j^f a_j + \sum_{m \in \text{inact}, m \neq i} k_m^f (1-a_j), \qquad (2)$$

where j is the index of effective kinases in the active state, and m is that in the inactive state.

The parameters, k_j^f and k_j^f, are constant for the normal condition, but reduced to smaller values under the drug (kinase inhibitor) conditions. We chose one of downstream variables (a_i) as the normalized neurite length.

We prepared 160 kinases and 167 drugs in total. The parameters and their reduction rates, k_j^f and k_j^f, by the drugs were randomly selected and we have 167 inhibition patterns of kinases and the corresponding neurite length by computing the reaction equations numerically. The highest 33 % of drugs in neurite length and the lowest 33 % were labelled one (elongation) and zero (shrinkage), respectively (Fig. 1B). Finally, we had the 112 x 160 matrix, Z, which has an element, $z_{i,j}$, representing the inhibition level of the j-th kinase by the i-th drug.

B. Classification algorithms

Ridged partial least squares (RPLS)

Partial least squares (PLS) is a tool for linear regression of continuous variables and a tool for dimension reduction [15], [17], [18], [19]. To apply PLS to binary classification problems, Fort and Lambert-Lacroix proposed a new method combining PLS and Ridge penalized logistic regression, which is called Ridge PLS (RPLS) [22]. The algorithm of RPLS is divided into the following two steps: the regularization step and the dimension-reduction step.

Step 1: Regularization step (Ridge logistic regression step)

Ridge penalized logistic regression is applied in this step and the estimators are used as target variables for the following PLS. In logit models, the conditional probability of Y given X is expressed by

$$P(\mathrm{Y}=1|\mathrm{X}=x;\gamma) = \frac{1}{1+\exp(\gamma x)}, \qquad (3)$$

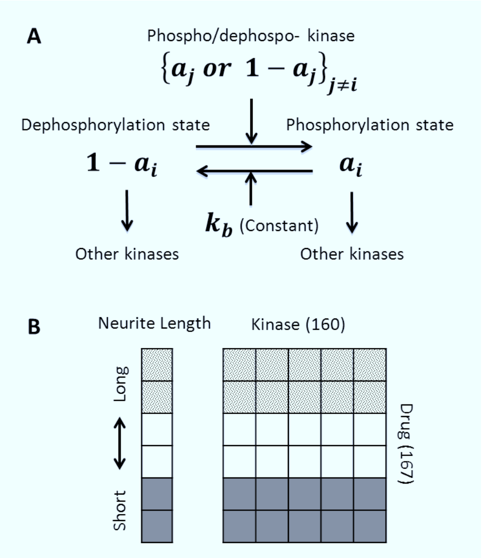

Fig. 1. (**A**) Schematic figure of biochemical reaction by the kinases. The jth kinase is activated when it is phosphorylated or dephosphorylated and then phosphorylates the ith kinase. (**B**) Kinase inhibitor rate matrix (160 kinases by 167 drugs). Neurite length vector was computed from the drug application in the corresponding row. Top 33% in the neurite length (dotted) were labeled as one (elongation) and bottom 33% (gray) were labeled as zero (shrinkage).

where γ is a weight vector and X is $(\mathbf{1}_n, z_{.,1}, \cdots, z_{.,\kappa})$ with the kinase inhibition level vector, $z_{.,i}$, and the n-dimentional vector, $\mathbf{1}_n$, whose elements are all one.

The ridge estimator is defined as the unique maximizer of the penalized cost function,

$$l^{Ridge}(\gamma) = \sum_{i=1}^{n}\{y_i\eta_i - \ln(1+\exp(\eta_i(\gamma)))\} + \frac{1}{2}\lambda\gamma^T\Sigma^2\gamma, \qquad (4)$$

where $\lambda > 0$ is the shrinkage parameter, and Σ^2 is a diagonal matrix with entries $\Sigma^2_{1,1} = 0$,

$$\Sigma^2_{j,j} = \sum_{i=1}^{n}\left(\mathrm{X}_{i,j} - \frac{1}{n}\mathbf{1}_n^T\mathrm{X}_{.,\mathrm{j}}\right)^2, \quad j \in \{2...p\}. \qquad (5)$$

The estimation is computed as the limit of a converging Newton-Raphson sequence; this algorithm is known as the RIRLS [23]. Let $W(\gamma)$ be the diagonal $n \times n$ matrix with diagonal entries $W_{i,i}(\gamma) = \pi_i(\gamma)(1-\pi_i(\gamma))$. Each iteration divides into two steps,

$$\begin{aligned} z^{(t)} &= \mathrm{X}\gamma^{(t)} + (W^{(t)})^{-1}(y-\pi^{(t)}), & (6)\\ \gamma^{(t+1)} &= (\mathrm{X}^{\mathrm{T}}W^{(t)}\mathrm{X} + \lambda\Sigma^2)^{-1}\mathrm{X}^TW^{(t)}z^{(t)}, & (7)\end{aligned}$$

where $W^{(t)}$ and $\pi^{(t)}$ are shorthand notations for $W(\gamma^{(t)})$ and $\pi(\gamma^{(t)})$. RIRLS can thus be considered as an iterative weighted least square regression of an $\mathcal{R}^n$-valued pseudo-variable $z^{(t)}$ onto the columns of X. We denote this algorithm by RIRLS (y, x).

Step 2: Dimension reduction step (Weighted PLS step)

PLS defines κ W-orthogonal scores $(t_k)_{1\leq k \leq \kappa}$, linear combinations of the columns of X such that for all k, $\mathbf{1}_n^T W_{t_\kappa}$=0 and (ii) performs a W-weighted least squares regression of y on $(\mathbf{1}_n, t_1, ...t_\kappa)$. This yields the decomposition

$$\begin{aligned} y &= q_0\mathbf{1}_n + q_1 t_1 +, ..., +q_\kappa t_\kappa + f_{\kappa+1} && (8)\\ &= \mathrm{X}\hat{\gamma}^{PLS,\kappa} + f_{\kappa+1}, && (9)\end{aligned}$$

where the residual term $f_{\kappa+1}$ is W-orthogonal to the vectors $(\mathbf{1}_n, t_1, ..., t\kappa)$. Contrary to classical dimension-reduction methods (such as PCR), the scores depend on the response vector y; roughly speaking, given $(t_k)_{1\leq k\leq l}$, t_{l+1} is the linear combination of the columns of X, i.e. is of the form $t_{l+1} = \mathrm{X}c$, which is the most informative on the residual response variable f_{l+1}, when information is defined in terms of the weighted covariance $|Cov(\sqrt{W}\mathrm{X}c, \sqrt{W}f_{l+1})|$ ($\sqrt{W}$ denotes the square root matrix of W)[14]. While the maximal number of PLS scores κ_{max} can be lower than rank(X). in practice, it is often equal to rank(X). Helland [14] shows that the weighted PLS regression applied with $\kappa = \kappa_{max}$ is nothing more than the weighted least squares regression. In the literature, PLS is usually derived with $W = I$, the identity matrix; we thus detail the algorithm in the weighted case. Let $\hat{\Sigma}$ be the $p \times p$ positive-definite diagonal matrix with diagonal entries $\Sigma_{j,j}$, $j \geq 2$, given by Equation (5).

1) $\mathrm{X}^s = \mathrm{X}\hat{\Sigma}^{-1}$, $t_0 = \mathbf{1}_n$, $E_0 = \mathrm{X}^s$; $f_0 = y$.
2) for $k = 0, \cdots, \kappa$

$$\begin{aligned} q_k &= t_k^T W f_k / (t_k^T W t_k), && (10)\\ f_{k+1} &= f_k - q_k t_k && (11)\\ E_{k+1} &= E_k - t_k t_k^T W E_k / (t_k^T W t_k), && (12)\\ t_{k+1} &= E_{k+1} E_{k+1}^T W f_{k+1}. && (13)\end{aligned}$$

Hereafter, this procedure is denoted by WPLS $(y, \mathrm{X}, W, \kappa)$. If X is full column-rank, this algorithm determines a unique estimate $\hat{\gamma}^{PLS,\kappa}$ satisfying $y - f_{k+1} = \mathrm{X}\hat{\gamma}^{PLS,\kappa}$; if X is not full column-rank, the procedure above yields the minimal norm vector among all the vectors verifying $y - f_{k+1} = \mathrm{X}\gamma$.

Predictiction of test dataset

In the RPLS, the estimator $\hat{\gamma}^{PLS}$ was defined by the following.

1) $(z^\infty, W^\infty) \leftarrow RIRLS(y, \mathrm{X}, \lambda)$
2) $\hat{\gamma}^{PLS,\kappa} \leftarrow WPLS(z^\infty, \mathrm{X}, W^\infty, \kappa)$

A detailed implementation is given in [22]. The first step builds a continuous response variable z^∞ for the input of PLS, the 'dispersion matrix' of which is $(W^\infty)^{-1}$. This explains the call, in the second step, to a weighted PLS procedure with weight W^∞. The use of X^s in WPLS and of Σ in the penalized ridge criterion makes our procedure invariant to the scaling of the data matrix.

For validation step, π_i was computed by using γ^{PLS} which was given by the training dataset. We labeled as 1 if $\pi_i > 0.5$, and 0 otherwise. We calculated the percentage of correct prediction from the predicted labels and those of the test dataset.

Other algorithms

We tested four other classifiers: a Naive Bayes classifier (NBC), a support vector classifier with a linear kernel (SVC(L)), a SVC with a gaussian kernel (SVC(GK)), and a classifier with random forest(RFC). We summarize key properties of those classifiers in the following:

- *Naive bayes classifier (NBC)*
  The Naive Bayes Classifier technique is based on the so-called Bayesian theorem and is particularly suited when the dimensionality of the inputs is high. Despite its simplicity, Naive Bayes can often outperform more sophisticated classification methods.
- *Support vector machine (SVM)*
  Support vector machine is a familiar method in the machine learning method and well known as a powerful classifier. In this study, we applied two types of the kernels: linear and gaussian.
- *Random forest classification (RFC)*
  Random forest classification is an algorithm for classification developed by Leo Breiman [9] that uses an ensemble of classification trees [10], [11], [12]. Each of the classification trees is built using a bootstrap sample of the data, and at each split the candidate set of variables is a random subset of the variables. Thus, random forest uses both bagging (bootstrap aggregation), a successful approach for combining unstable learners [12],[13], and random variable selection for tree building. Each tree is unpruned (grown fully), so as to obtain low-bias trees; at the same time, bagging and random variable selection result in low correlation of the individual trees. The algorithm yields an ensemble that can achieve both low bias and low variance (from averaging over a large ensemble of low-bias, high-variance but low correlation trees).

Using the kinase inhibiter matrix (Fig. 1B), we trained five statistical classifiers including RPLS and performed prediction test for the untrained drug data. As a classification test, each classifier was trained to predict those label from the corresponding rows, and the percentage of correct classification was calculated with cross-validation. In the cross-validation, we divided those rows into 10 groups. One group was used as test data and the rest was used to train a classifier. This procedure was repeated until all group was used as test data.

III. Results

The percentages of correct predictions of the five algorithms for four datasets are shown in Fig. 2. For the prediction by RPLS, the parameters, λ and κ, were decided such that the score took the maximum for the dataset 1. The same parameters were used for the prediction of the other dataset. Nevertheless, RPLS showed good performances for all the datasets. It was not the best classifier for the dataset 4, but was still superior to SVM. This result suggests that the dimension of the dataset could be well reduced in the PLS algorithm, as we expected.

IV. Conclusion

We applied the method of RPLS to the synthetic datasets to find key kinases which contribute to neurite elongation. By comparing with the other classifiers, we showed that RPLS can present the best performance, as we expected. This suggests that RPLS is an effective algorithm for biological data which has a correlation among target elements. The application of RPLS to dataset given by biological experiments is the future work.

ACKNOWLEDGMENT

The authors would like to thank S. Kozawa for helpful discussions. This work was supported by JSPS KAKENHI Grant Number 11J08024 and MEXT KAKENHI Grant Number 23127509.

References

[1] Blenis J, Signal transduction via the MAP kinases: proceed at your own RSK, Proc. Natl. Acad. Sci. USA, 90, 5889-5892, 1993.

[2] Kumar S, Boehm J, Lee JC, p38 MAP kinases: key signalling molecules as therapeutic targets for inflammatory diseases, Nat. Rev. Drug Discovery, 2, 717-726, 2003.

[3] Buchser WJ, Slepak TI, Gutierrez-Arenas O, Bixby JL, Lemmon VP, Kinase/phosphatase overexpression reveals pathways regulating hippocampal neuron morphology, Mol. Syst. Biol. 6:391, 2010.

[4] Patterson SL, Abel T, Deuel TAS, Martin KC, Rose JC, Kandel ER, Recombinant BDNF Rescues Deficits in Basal Synaptic Transmission and Hippocampal LTP in BDNF Knockout Mice, Neuron, 16, 1137-1145, 1996.

[5] Cafferty WBJ, Gardiner NJ, Das P, Qiu J, McMahon SB, Thompson SWN, Conditioning Injury-Induced Spinal Axon Regeneration Fails in Interleukin-6 Knock-Out Mice, J. Neurosci., 24, 4432-4443, 2004.

[6] http://www.broadinstitute.org/cmap/

[7] Yang X, Xie L, Li Y, Wei C, More than 9,000,000 Unique Genes in Human Gut Bacterial Community: Estimating Gene Numbers Inside a Human Body, PLoS ONE, 4(6): e6074, 2009.

[8] Johnstone AL, Reierson GW, Smith RP, Goldberg JL, Lemmon VP, Bixby JL: A chemical genetic approach identifies piperazine antipsychotics as promoters of CNS neurite growth on inhibitory substrates, Mol. Cell. Neurosci., 50:125-135. 2012.

[9] Breiman L: Random forests. Machine Learning, 45:5-32. 2001.

[10] Breiman L, Friedman J, Olshen R, Stone C: Classification and regression trees. New York: Chapman & Hall; 1984.

[11] Ripley BD: Pattern recognition and neural networks. Cambridge: Cambridge University Press; 1996.

[12] Hastie T, Tibshirani R, Friedman J: The elements of statistical learning. New York: Springer; 2001.

[13] Breiman L: Bagging predictors. Machine Learning, 24:123-140. 1996.

[14] Helland, I. On the structure of partial least squares regression. Commun. Statist., Simulation Comput. 17581-17607, 1988.

[15] Wold, H. Soft modelling by latent variables: the non-linear iterative partial least squares (NIPALS) approach. In Gani, J. (Ed.). Perspectives in Probability and Statistics, Papers in Honour of M. S. Bartlett , London Academic Press, 117-142 , 1975.

[16] Daz-Uriarte R and de Andrs SA, Gene selection and classification of microarray data using random forest, BMC Bioinformatics, 7, 3-, 2006.

[17] Naes, T. and Martens, H. Comparison of prediction methods for multicollinear data. Commun. Statist. Simulation Comput. 14545-14576, 1985.

[18] Helland, I. On the structure of partial least squares regression. Commun. Statist., Simulation Comput. 17581-17607, 1988.

[19] Nguyen, D. and Rocke, D. Tumor classification by partial least squares using microarray gene expression data, Bioinformatics, 1839-1850, 2002

[20] Massy, WF, Principal components regression in exploratory statistical research. J. Amer. Statist. Assoc. 60234-60246, 1965.

[21] Frank I. and Friedman J. A statistical view of some chemometrics regression tools, with discussion. Technometrics 35, 109-148, 1993.

[22] Fort G. and Lambert-Lacroix S. Classification using partial least squares with penalized logistic regression, Bioinformatics, 21, 1104-1111, 2005.

[23] Green.P. Iteratively reweighted least squares fo maximum likelihood estimation a nd some robust and resistant alternatives, J.R.Statist.Soc. B, 46(2), 149-192. 1984.

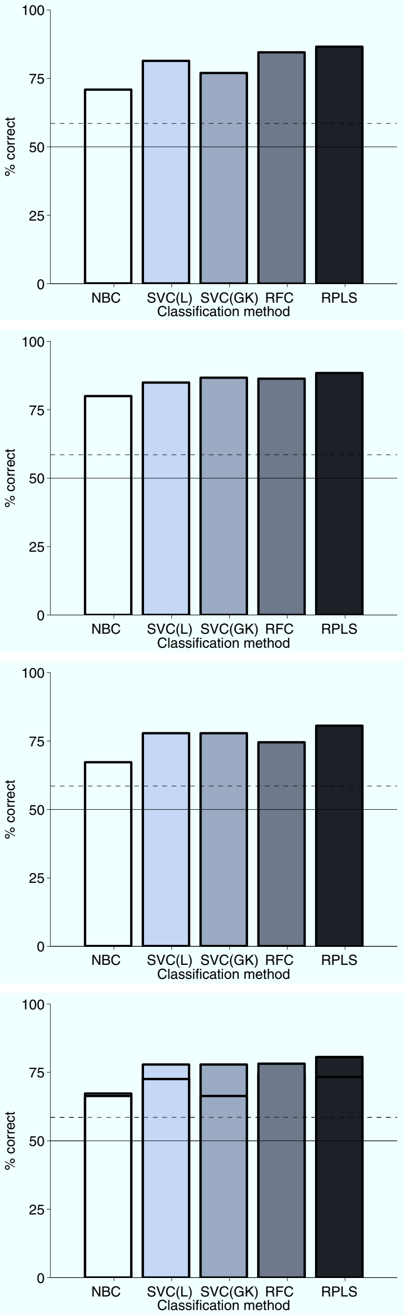

Fig. 2. Percentages of correct predictions for five applied classifiers; naive bayes classifier (NBC), support vector classifier with a linear kernel (SVC(L)), SVC with a gaussian kernel (SVC(GK)), classifier with random forest(RFC), and ridge partial least squares (RPLS). The results given from four dataset are shown. The horizontal solid line represents 95% significant of the binomial test, and the dashed line indicates the chance level.