Adaptive Locality-Effective Kernel Machine for Protein Phosphorylation Site Prediction

Paul D. Yoo, Yung Shwen Ho, Bing Bing Zhou, Albert Y. Zomaya
Advanced Networks Research Group
School of Information Technologies (J12)
University of Sydney, NSW 2006, Australia
dyoo4334@it.usyd.edu.au

Abstract

In this study, we propose a new machine learning model namely, Adaptive Locality-Effective Kernel Machine (Adaptive-LEKM) for protein phosphorylation site prediction. Adaptive-LEKM proves to be more accurate and exhibits a much stable predictive performance over the existing machine learning models. Adaptive-LEKM is trained using Position Specific Scoring Matrix (PSSM) to detect possible protein phosphorylation sites for a target sequence. The performance of the proposed model was compared to seven existing different machine learning models on newly proposed PS-Benchmark_1 dataset in terms of accuracy, sensitivity, specificity and correlation coefficient. Adaptive-LEKM showed better predictive performance with 82.3% accuracy, 80.1% sensitivity, 84.5% specificity and 0.65 correlation-coefficient than contemporary machine learning models.

1. Introduction

Post-translational modifications are observed on almost all proteins analysed to date. These modifications have a substantial influence on the structure and functions of proteins. Phosphorylation at the serine, threonine and tyrosine residues by enzymes of the kinase and phosphatise super-families is one of the most frequent forms of post-translational modifications in intracellular proteins. Phosphorylation has a significant impact on diverse cellular signalling processes. The phosphorylation-dependence signals function diversely in animals such as a regulation of differentiation of cells, a trigger of progression of the cell cycle, and a control of metabolism, transcription, apoptosis, cytoskeletal rearrangements and so forth [6] [12] [15] [17] [20] [25] [31].

During phosphorylation, a phosphate molecule is placed on another molecule resulting in the functional activation or inactivation of the receiving molecule. The phosphorylation of any site on a given protein can also alter the function or localisation of that protein. Phosphorylation of a protein is considered as a key event in many signal transduction pathways of biological systems [5]. It is thus important for us to be able to accurately determine the phosphorylation state of proteins so as to better identify the state of a cell.

In order to determine phosphoproteins and individual phosphorylation sites, various experimental tools have been used. However, many indicated that in vivo or in vitro identification of phosphorylation sites is labour-intensive, time-consuming and often limited to the availability and optimisation of enzymatic reactions [5] [31] [33]. Several large-scale phosphoproteomic data using the mass spectrometry approach have been collected and published [2] [3] [9]. These however are sill unfavourable in distinguishing the kinase-specific sites on the substrates. For an example, mass spectrometry methods have been shown to be disfavour in the identification of phosphate-modified residues, leading to underestimation of the extent of phosphorylation presents in vivo [23].

Due to the practical limitations of the afore-mentioned methods, many scientists now turn to computer-based methods. These methods not only efficiently handle massive amounts of protein data but also determine phosphoprotiens and identify individual phosphorylation sites from one dimensional atomic coordinates with high precision. Several computer-simulated machine learning techniques such as Artificial Neural Networks (ANNs) and Support Vector Machines (SVMs) have been extensively used in various biological sequence analyses and phosphorylation site prediction. These methods are built based on the assumption that neighbouring

978-1-4244-1694-3/08/$25.00 ©2008 IEEE

residues to the phosphorylated site represents the main determinant for kinase specificity [29] [33].

Although a large number of machine learning based methods were proved to be effective in the prediction of phosphorylation site, several important issues that can potentially degrade the performance of machine learning or statistical-based methods have been largely ignored. It has been widely recognised that the high dimensionality of protein sequence data not only causes a dynamic increase in computational complexity but also induces into the overfitting/generalisation problem of non-parametric methods. With machine learning models, better generalisation and faster training (computationally efficient) can be achieved when they have fewer weights to be adjusted by fewer inputs.

This study aims to develop an accurate and stable machine learning model for phosphorylation site prediction. Our proposed model called, Adaptive Locality-Effective Kernel Machine (Adaptive-LEKM) uses a semi-parametric form of the existing support vector machine. In addition, with the boosting algorithm, it adaptively combines the learners to find an optimised fit for its given phosphoprotiens. In our experiments, the Adaptive-LEKM excels in efficiently processing high dimensional protein data with a much more accurate and stable predictive performance over existing models. The novel features of this study are the use of a new machine learning based semi-parametric model, and the use of unique training dataset (PS-Benchmark_1) contains experimentally verified phosphorylation sites manually extracted from major protein sequence databases and the literature.

2. Materials and methods

2.1. PS-Benchmark_1 dataset

The fair comparison and assessment of each model is complicated as all use different phosphorylation site datasets in the literature. In this study, we use a newly developed comprehensive dataset, namely PS-Benchmark_1 for the purpose of benchmarking sequence-based phosphorylation site prediction methods. It is widely known that accurate classification is highly dependent upon high quality data sets of both positive and negative examples. However, such a golden standard datasets are not yet available for protein phosphorylation site prediction. PS-Benchmark_1 contains experimentally verified phosphorylation sites manually extracted from major protein sequence databases and the literature. The dataset comprises of 1,668 polypeptide chains and the chains are categorised in four major kinase groups, namely cAMP-dependent protein kinase/protein kinase G/protein kinase C extended family (AGC), calcium/calmodulin-dependent kinase (CAMK), cyclin-dependent kinase-like kinase (CMGC) and tyrosine kinase (TK) groups. The dataset comprises of 513 AGC chains, 151 CAMK chains, 330 CMGC chains, and 216 TK chains. The dataset is non-redundant in a structural sense: each combination of topologies occurs only once per dataset. Sequences of protein chains are taken from the Protein Data Bank (PDB) [4], Swiss-Prot [1], Phospho3D [33], Phospho.ELM [7] and literature.

2.2. Proposed model

Protein sequence data can be mathematically viewed as points in a high dimensional space. For example, a sequence of 10 amino acids represents a search space of 20^{10} possibilities and requires a network of 200 inputs. In many applications, the curse of dimensionality is one of the major problems that arise when using non-parametric techniques [13]. Learning in the high dimensional space causes several important problems. First, the good data fitting capacity of the flexible "model-free" approach often tends to fit the training data very well and thus, have a low bias. However, the potential risk is the overfitting that causes high variance in generalisation. In general, the variance is shown to be a more important factor than the learning bias in poor prediction performance [8]. Second, with the high dimensional data, as the number of hidden nodes of the network is severely increased, it eventually leads to an exponential rise in computational complexity. A high complexity model generally shows a low bias but a high variance [21]. On the other hand, a model with low complexity shows a high bias but a low variance. Hence, a good model balances well between model bias and model variance. This problem is generally regarded as the term "*bias and variance tradeoff*".

One solution to the problems above can be semi-parametric modelling. Semi-parametric models take assumptions that are stronger than those of non-parametric models, but are less restrictive than those of parametric model. In particular, they avoid most serious practical disadvantages of non-parametric methods but at the price of an increased risk of specification errors. The proposed model, Adaptive-LEKM takes a form of the semi-parametric model and it finds the optimal trade-off between parametric and non-parametric models. So, it can have advantages of both models while effectively avoiding the curse of dimensionality. The Adaptive-LEKM contains the evolutionary information represented with the local model. Its global model works as a collaborative filter

that transfers the knowledge amongst the local models in formats of the hyper-parameters. The local model contains an efficient vector quantisation method.

Vector Quantisation (VQ) is a lossy data compression technique based on the principle of book coding. Its basic idea is to replace with key values from an original multidimensional vector space into values from a discrete subspace of lower dimension. The lower-space vector requires less storage space and the data is thus compressed. Consider a training sequence consisting of M source vectors, $T=\{x_1, x_2, ..., x_m\}$. M is assumed to be sufficiently large and so that all the statistical properties of the source are captured by the training sequence. We assume that the source vectors are k-dimensional, $X_m=(x_{m,1}, x_{m,2}, ..., x_{m,k})$, $m=1,2,...,M$. These vectors are compressed by choosing the nearest matching vectors and form a codebook consisting the set of all the codevectors. N is the number of codevectors, $C=\{c_1,c_2,...,c_n\}$ and each codevector is k-dimensional, $c_n=(c_{n,1},c_{n,2},...,c_{n,k})$, $n=1,2,...,N$.

S_n is the nearest-neighbour region associated with codevector c_n, and the partitions of the whole region are denoted by $P=\{S_1,S_2,...,S_N\}$. If the source vector X_m is in the region S_n, its approximation can be denoted by $Q(X_m)=c_n$, if $X_m \in S_n$. The *Voronoi* region is defined by:

$$V_i = \{x \in R^k : \|x - c_i\| \le \|x - c_j\|, \text{for all } j \ne i\}$$

The centroid condition requires the codevector c_n should be average of all those training vectors that are in its *Voronoi Region* S_n

$$c_n = \frac{\sum_{Xm \in Sn} X_m}{\sum_{Xm \in Sn} 1}, n = 1,2,..., N$$

As a key collaborator, we use one effective kernel method to construct the global model so-called Support Vector Machine (SVM). The SVM used in the Adaptive-LEKM is the modified version of SVM^{light} package with an RBF kernel for the classifiers. The hyperparameters used in the SVM were optimised using a 7-fold cross-validation. In order to find optimal values for the hyperparameters, a number of values were considered and tested against the newly built PS-Benchmark_1 dataset and the optimal values were chosen are C: 1.0, γ: 0.04, and ε: 0.1.

In the literature, it is claimed that one of the most serious problems with SVMs is the high algorithmic complexity and extensive memory requirements of the required quadratic programming in large-scale tasks. As observed in above equation, SVM extracts worst-case examples x_i and use statistical analysis to build large margin classifiers. However, in Adaptive-LEKM, we use the centroid vector of each *voronoi* region which can be expressed as:

$$Q_i(X_m) = c_i = \frac{\sum_{Xm \in Si} x_m}{N}, i = 1,2,..., N$$

To construct a semi-parametric model, we substitute $Q_i(X)$ for each training sample x_i used in the SVM decision function. Given a d-dimensional input vector, $x_i = (x_1, x_2, ..., x_d)$ with two labels, $y_i \in \{+1, -1\}$ (i =1, 2, ..., N), the Adaptive-LEKM's approximation can be written as:

$$f(x) = \text{sgn}\left(w \cdot \phi(x) + b\right) = \text{sgn}\left(\sum_{i=1}^{\ell} y_i a_i k(x, c_i) + b\right)$$

and the following quadratic program:

$$\text{maximise } W(a) = \sum_{i=1}^{\ell} a_i - \frac{1}{2}\sum_{i,j=1}^{\ell} a_i a_j y_i y_j k(Q_i(x), Q_j(x))$$

subject to $a_i \ge 0$, $i = 1,..., \ell$, and $\sum_{i=1}^{\ell} a_i y_i = 0.$

where ℓ is the number of training patters; a_i are the parameters of the SVM; $k(\cdot,\cdot)$ is a suitable kernel function, and b is the bias term.

The SVM is considered as a purely non-parametric model, whereas the Adaptive-LEKM can be considered as semi-parametric model as it adopts the method of grouping of the associated input vectors in each class i. Hence, the performance of proposed model has some advantages in comparison to the pure parametric models and pure non-parametric models in terms of learning bias and generalisation variance especially on high dimensional protein datasets.

As the Adaptive-LEKM uses the centroid vector of each nearest neighbour region, we can obtain the optimal representations by finding right size of each sub regions. In other words, a good trade-off between parametric and non-parametric can be found by adjusting the size of each sub-region. If the feature space is partitioned to too many sub-regions, the model becomes closer to non-parametric model. So, it is eventually susceptible to overfitting which causes high model variance problem. Contrarily, if the space is divided into too small number of regions, the codevectors cannot correctly represent the original dataset (as they miss too much information). And the model eventually produces high leaning bias. Hence, it is crucial to find a good trade-off between the parametric and non-parametric models.

In order to maximise the performance of the Adaptive-LEKM, we utilised a network boosting method called Adaptive Boosting (AdaBoost). In general, boosting is known as a technique to improve the performance of any base machine learning algorithms. The AdaBoost algorithm was proposed by Freund and Schapire [10] and it was shown to be a solution for many practical difficulties of previous boosting algorithms. Boosting combines weak learners to find a highly accurate classifier or better fit for the training set [28]. In this study, the AdaBoost was modified for the LEKM for the network boosting. As observed in our experiments, the modified AdaBoost was tested with the LEKM and showed that it can fit into its architecture for more accurate prediction of phosphorylation sites. A standard boosting algorithm can be written as:

Given: $(x_1, y_1),\ldots,(x_{NV}, y_{NV})$ where $x_i \in X, y_i \in Y = \{-1,+1\}$

Initialise $D_1(i) = 1/NV$

For $t = 1,\ldots,T$:

- Find the classifier $h_t : X \to \{-1,+1\}$ that minimises the error with respect to the distribution D_t

  $h_t = \arg\min_{h_j \in H} \varepsilon_j$, where $\varepsilon_j = \sum_{i=1}^{NV} D_t(i)[y_i \neq h_j(x_i)]$
- Get weak hypothesis $h_t : X \to [0,1]$
- Choose $a_t \in R$, typically $a_t = \frac{1}{2} in \frac{1-\varepsilon_t}{\varepsilon_t}$ where ε_t is the weighted error rate of classifier h_t
- Update:

$$D_{t+1}(i) = \frac{D_t(i)\exp(-a_t y_i h_t(x_i))}{Z_t}$$

where

NV = total number of training vectors,

X = a domain or instance space of each x_i belong to,

Y = a label set of each label y_i,

Z_t = a normalisation factor (chosen so that D_{t+1} will be a distribution),

R = its sign is the predicted label {-1,+1}.

Output the final hypothesis:

$$H(x) = sign\left(\sum_{t=1}^{T} a_t h_t(x)\right)$$

In the training se, each x_i belong to a domain X, and each label y_i is in a label set Y. Here, the Y should be {-1,+1} as phosphorylation sites are indicated as positive (+1) or negative (-1) values only. After selecting an optimal classifier h_t for the distribution D_t, the examples x_i that the classifier h_t identified correctly are weighted less and those that it identified incorrectly are weighted more. Therefore, when the algorithm is testing the classifiers on the distribution $D_t + 1$, it will select a classifier that better identifies those examples that the previous classifier missed. At each iteration, the AdaBoost embedded in Adaptive-LEKM constructs weak learners based on this method called weighted examples.

2.3. Training, testing, and validation

For the fair comparison of our proposed model, we adopted a seven fold cross-validation scheme for the model evaluation. Random dataset selection and testing was conducted seven times for each different window size dataset. When multiple random training and testing experiments were performed, a model was formed from each training sample. The estimated prediction accuracy is the average of the prediction accuracy for the models and each window size, derived from the independently and randomly generated test divisions. We used the window size of 9 for tyrosine and threonine, and 11 for serine sites [29]. A window size of 9 means 19 amino acids with the tyrosine, threonine or serine site is located at the centre of the window.

The performance of Adaptive-LEKM is measured by the accuracy (Ac: the proportion of true-positive and true-negative residues with respect to the total positives and negatives residues), the sensitivity (Sn: the proportion of correctly predicted phosphorylation site residues with respect to the total positively identified residues), the specificity (Sp: the proportion of incorrectly predicted site residues with respect to the total number of phosphorylation site residues) and correlation coefficient (Cc: It balances positive predictions equally with negative predictions and varies between -1 and 1.). Cc reflects a situation in that a method which predicts every residue to be positive, shows prediction accuracy of 100% in detecting positive sites, however 0% accuracy for negative residues. Hence, high value of Cc means that the model is regarded as a more robust prediction system. In addition to the four measures above, the performance of each model is additionally measured by Type I and Type II Error rates as incorrectly predicted residues can be as valuable as the correctly predicted residues for further modification of the model. Type I Error means experimentally verified unmodified sites that are predicted (incorrectly) to be modified; And Type II Error indicates experimentally verified modified sites that are predicted (incorrectly) to be unmodified. The Sn, Sp, Ac and CC can be expressed in terms of true positive (TP), false negative (FN), true negative (TN) and false positive (FP) predictions.

$$Sn = \frac{TP}{TP + FN'}, \qquad Sp = \frac{TN}{TN + FP'}$$

and,

$$Cc = \frac{(TP \times TN) - (FN \times FP)}{\sqrt{(TP + FN) \times (TN + FP) \times (TP + FP) \times (TN + FN)}}$$

3. Experimental results

Our experiments consist of two consecutive steps. First the predictive performance of our proposed machine learning model, Adaptive Locality-Effective Kernel Machine (Adaptive-LEKM), specially designed for the high dimensional problem of protein sequence data is compared with other seven contemporary machine learning models in terms of prediction accuracy, sensitivity, specificity, correlation-coefficient, type I and type II errors on newly built PS-Benchmark_1 dataset. Second, to provide more in-depth and analytic results of our proposed model, Adaptive-LEKM is tested on each four major kinase families and we compare its results with the consensus results of the literature.

3.1. Comparison with other machine learning models

The predictive performance of our proposed model (Adaptive-LEKM) was compared with seven other existing state-of-the-art machine learning models such as General Regression Neural Network (GRNN), Radial Basis Neural Network (RBFN), Multi-Layered Perceptron (MLP), kernel Nearest Neighbor (kNN), Decision Tree (J48), kernel Logistic Regression (KLR), and two different transductive Support Vector Machines, namely Support Vector Machine (SVM) and Locality-Effective Kernel Machine (LEKM). Table 1 shows evaluation results of each model in terms of Accuracy (Ac), Sensitivity (Sn), Specificity (Sp), Correlation-Coefficient (Cc), Variance (Var) and Time on PS-Benchmark_1 dataset.

As shown in Table 1, one of our models, LEKM was shown to be successful as it reaches the best model stabilisation (Var: 0.020) and less computational requirements (Time: 22.421). However, one of the methods used in LEKM, semi-parametric approximation brought into a slightly less accurate learning. Hence, we utilised AdaBoost algorithm for the fine tuning of the LEKM and it (Adaptive-LEKM) finally achieved the best accuracy with a fair level of model stableness and reduced complexity. In addition, Adaptive-LEKM achieved much better model robustness than other methods with the Cc of 0.646. Our methods used in Adaptive-LEKM, semi-parametric approximation and adaptive tuning of the model using AdaBoost were confirmed to be more suitable in processing high dimensional protein data than pure non-parametric approaches.

Table 1. Prediction results of machine learning models on PS-Benchmark_1 dataset.

Models	Ac	Sn	Sp	Cc	Var.	Time
Ada-LEKM	**0.823**	**0.801**	**0.845**	**0.646**	**0.022**	**35.745**
SVM	0.798	0.787	0.809	0.596	0.030	32.886
LEKM	**0.783**	**0.773**	**0.792**	**0.565**	**0.020**	**22.421**
kNN	0.767	0.753	0.781	0.534	0.032	35.630
GRNN	0.759	0.724	0.793	0.518	0.041	85.422
MLP	0.752	0.715	0.789	0.505	0.046	180.344
RBFN	0.737	0.685	0.788	0.475	0.044	68.654
DT (J48)	0.732	0.718	0.747	0.465	0.025	8.393
KLR	0.726	0.682	0.772	0.456	0.038	156.690

Figure 1 shows that the comparison of prediction scores simulated by the Adaptive-LEKM and the original SVM on a protein chain, Swiss-Prot Entry: O75553. The protein chain has 588 residues with two tyrosine and one serine sites at the residue 198, 220 and 524 respectively. As shown in Figure 1, SVM's signal at the site is generally only around at 0.3 point with many fluctuating neighbouring signals so that the site may hardly be distinguished. On the other hand, Adaptive-LEKM provides very clear indication of the phosphorylation site at the residue 524 and its signal is generally stronger than that of other methods by reaching almost 0.4 point (0.37921). Adaptive-LEKM offers an additional level of advantages over other machine learners with more clear and strong indication of site locations.

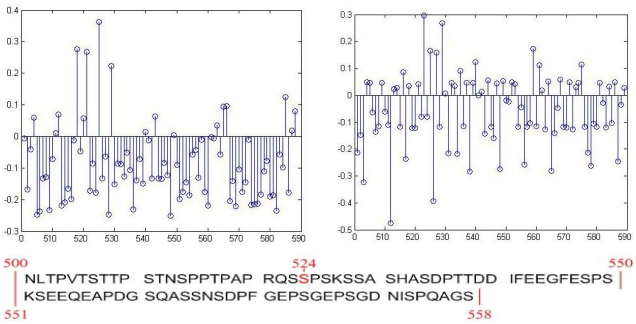

Figure 1. Prediction scores simulated by Adaptive-LEKM and SVM.

3.2. Predictive performance of Adaptive-LEKM on major kinase families

Now, we look at the experimental results obtained by Adaptive-LEKM on four main kinase families in terms of Ac, Sn, Sp, Cc, Type I ER and Type II ER. Table 2 compares the results of Adaptive-LEKM with the consensus results of the literature. In general, Adaptive-LEKM showed about 9% better prediction

accuracy than the consensus results. As for the model stableness, Adaptive-LEKM also achieved a fairly low level of average variance in four evaluation measures. The sensitivity of Adatpive-LEKM on CDK, PKA and PKC kinase families are distinguishably higher than the consensus results. It means that the more stable prediction capability of our model comes with effectively reducing the false negative values (Type I ER). Type I ER indicates experimentally verified unmodified sites that are predicted (incorrectly) to be modified.

Table 2. Prediction results of Adaptive-LEKM for the four kinase families.

K-Families	Ac	Sn	Sp	Cc	Type I ER	Type II ER
CDK	**0.909**	**0.895**	**0.921**	**0.817**	**0.043**	**0.046**
	0.777	0.455	0.992	0.900		
CK2	**0.918**	**0.881**	**0.948**	**0.835**	**0.029**	**0.051**
	0.840	0.765	0.888	0.660		
PKA	**0.891**	**0.843**	**0.929**	**0.779**	**0.039**	**0.069**
	0.816	0.561	0.987	0.640		
PKC	**0.827**	**0.731**	**0.903**	**0.650**	**0.053**	**0.118**
	0.726	0.475	0.898	0.420		
Avg.	**0.886**	**0.838**	**0.925**	**0.770**	**0.041**	**0.071**
	0.790	0.564	0.941	0.655		
Var.	**0.041**	**0.074**	**0.019**	**0.083**	**0.010**	**0.032**
	0.050	0.142	0.056	0.196		

The experimental results of Adaptive-LEKM are written in bold and others are the consensus results of literature obtained by [18].

4. Discussion

Over the past decades, many computational prediction algorithms have been developed for various proteomic studies. They have evolved from simple linear statistics to complex machine learners. However, the most significant breakthroughs were the incorporation of new biological information into an efficient prediction model and the development of new models which can efficiently exploit suitable information from its primary sequence. For example, the exploitation of evolutionary information that is available from protein families has brought significant improvements in the prediction of protein secondary structure (about 6-8%) [11] [19] [26] [27] [34].

Compared to protein structure predictions, a feeble effort to find suitable information/representations for phosphorylation site prediction has been reported. Like used in protein secondary structure prediction, mostly uses the evolutionary information in the format of Position Specific Scoring Matrix (PSSM or sequence profile) [14] [16] [24] [32]. The behind theory of using sequence profile is based on the fact that the sequence alignment of homologous proteins accords with their structural alignment and aligned residues usually have similar structures. Thus, the sequence profile can provide more information about structure than single sequence to its learner.

Although the sequence profile provides more structural information, the structural information resides in sequence profile may not be a significant importance in the case of phosphorylation site prediction. It has been observed that approximately only ten neighbouring residues are the major determinants of phosphorylation sites. Many models have been built on this observation and performed reasonably well with a number of specific kinases. However, the specificity determinants and rules remain elusive for a large number of protein kinases that display a number of substrates sharing little or no sequence similarity in the known phosphopeptides [33]. Furthermore, most databases searched by current alignment tools like PSI-BLAST not only contains a number of non-phosphoprotiens, but also generates a large number of irrelevant hits in the protein databases [5].

The encoding methods discussed above are employed by most well-known protein structure predictors and were shown to be useful as they sufficiently contain information required for general protein structure prediction tasks. However, as for the phosphorylation site prediction, as it not only involves various chemical interactions but is known as a non-structural prediction task, the encoding method like PSSM may not be suitable for this problem. In the literature, encoding scheme proposed for phosphorylation site prediction is far less than ones for other proteomic applications. As discussed above, existing methods shows have several critical drawbacks for phosphorylation site prediction. Hence, we emphasise that researchers should devote their effort to seeking a suitable representation of amino acids for phosphorylation site prediction to reach the upper boundary of prediction accuracy.

5. Conclusion

This paper identified the effectiveness and utility of the newly proposed machine learning model, namely Adatpvie-LEKM for phosphorylation site prediction. This study addressed two important issues in protein phosphorylation site research. First, for a given set of high dimensional protein data, the combination of a parametric local model with a non-parametric global model provided a way of fine-tuning the model by the adjustment of a single smoothing parameter σ as well as providing efficient semi-parametric approximation.

This was demonstrated by the above experiments. The semi-parametric approach used in Adaptive-LEKM was shown to be effective by finding an optimal trade-off between parametric and non-parametric models with significantly reduced computations. With the newly built PS-Benchmark_1 dataset, Adaptive-LEKM achieved the best prediction accuracy when compared with the existing state-of-the-art machine learning models.

10. References

[1] Amos, B. Serendipity in bioinformatics, the tribulations of a Swiss bioinformatician through exciting times!. Bioinformatics 2000, 16: 48-64

[2] Ballif, B., A, Villen, J., Beausoleil, S.A., Schwartz, D,. and Gygi, S.P. Phosphoproteomic analysis of the developing mouse brain. Mol. Cell. Proteomics 2004, 3, 1093-1101

[3] Beausoleil, S.A, Jedrychowski, M, Schwartz, D, Elias, J.E., Villen, J., Li, J., Cohn, M.A., Cantley, L.C. and Gygi, S.P. Large-scale characterization of HeLa cell nuclear phosphoproteins. Proc. Natl Acad. Sci 2004, 101: 12130-12135

[4] Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindylov, I.N., Bourne, P.E. The Protein Data Bank. Nucleic Acids Research 2000, 28:235-242

[5] Blom, N., Sicheritz-Pontén, T., Gupta, R., Gammeltoft, S. and Brunak, S. Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence. Proteomics 2004, 4:1633-1649

[6] Cohen, P. The origins of protein phosphorylation. Nat Cell Biol 2002, 4(5):E127-E130

[7] Diella, F., Cameron, S., Gemund, C., Linding, R., Via, A., Kuster, B., Sicheritz-Ponten, T., Blom, N., Gibson, T.J. Phospho.ELM: a database of experimentally verified phosphorylation sites in eukaryotic proteins. BMC Bioinformatics 2005, 5:79

[8] Dietterich, T.G., Bakiri, G. Machine Learning bias, statistical bias and statistical variance of decision tree algorithms. Dept. Comput. Sci., Oregon State Univ., Corvallies, Tech. Rep., 1995.

[9] Ficarro, S.B., McCleland, M.L., Stukenberg, P.T., Burke, D.J., Ross, M.M., Shabanowitz, J., Hunt, D.F., and White, F.M. Phosphoproteome analysis by mass spectrometry and its application to Saccharomyces cerevisiae. Nat. Biotechnol 2002, 20: 301-305

[10] Freund, Y. and Schapire, R.E. Experiments with a new boosting algorithm. In Machine Learning: Proceedings of the Thirteenth International Conference 1996, 148-156

[11] Frishman, D. and Argos, P. Incorporation of non-local interactions in protein secondary structure prediction from the amino acid sequence. Protein Engineering 1996, 9: 133-142

[12] Graves, L., Bornfeldt, K. and Krebs, E. Historical perspectives and new insights involving the MAP kinase cascades. Advan. Sec. Mess. Phos. Res 1997, 31: 49-62

[13] Hardle, W., Muller, M., Sperlich, S., Warwatz, A. Nonparametric and Semiparametric models. Springer, New York 2004

[14] Hjerrild, M., Stensballe, A., Rasmussen, T.E., Kofoed, C.B., Blom, N., Sicheritz-Ponten, T., Larsen, M.R., Brunak, S., Jensen, O.N., and Gammeltoft, S. Identification of Phosphorylation Sites in Protein Kinase A Substrates Using Artificial Neural Networks and Mass Spectrometry. Proteome Research 2004, 3: 426-433

[15] Hunter, T. Signaling-2000 and beyond. Cell 2000, 10(1):113-127

[16] Iakoucheva, L.M., Radivojac, P., Brown, C.J., O'Connor, T.R., Sikes, J.G., Obradovic, Z. and Dunker, A.K. The importance of intrinsic disorder for protein phosphorylation. Nucleic Acids Research 2004, 32(3):1037-1049

[17] Johnson, L., Lowe, E., Noble, M., and Owen, D. The eleventh datta lecture. the structural basis for substrate recognition and control by protein kinases. FEBS Letters 1998, 430: 1-11

[18] Kim, J.H., Lee, J., Oh, B., Kimm, K., and Koh, I. Prediction of phosphorylation sites using SVMs. Bioinformatics 2004, 20(1):3179-3184

[19] King, R.D., and Sternberg, M.J.E. Identification and application of the concepts important for accurate and reliable protein secondary structure prediction. Protein Science 1996, 5: 2298-2310

[20] Koliba, K., and Druker, B. Protein tyrosine kinases and cancer. Biochim. Biophys. Acta 1997, 1333: F217-F248

[21] Larose, D.T. Discovering Knowledge in Data. Wiley, 2005

[22] Liu, J., Rost, B. Sequence-based prediction of protein domains. Nucleic Acids Research 2004, 32(12):3522-3530

[23] Mann, M., Ong, S.E., Gronborg, M., Steen, H., Jensen, O.N., Pandey, A. Analysis of protein phosphorylation using mass spectrometry: deciphering the phosphoproteome. Trends Biotechnol. 2002, 20(6):261-268

[24] Obenauer, J.C., Cantley, L.C., and Yaffe, M.B. Scansite 2.0: proteome-wide prediction of cell signaling interactions using short sequence motifs. Nucleic Acids Research 2003, 31(13):3635-3641

[25] Pinna, L.A. and Ruzzene, M. How do protein kinases recognize their substrates? Biochim. Biophys. Acta 1996, 1314: 191-225

[26] Rost, B., and Sander, C. Prediction of protein secondary structure at better than 70% accuracy. Journal of Molecular Biology 1993, 232: 584-599.

[27] Salamov, A.A., and Solovyev, V.V. Prediction of protein secondary structure by combining nearest- neighbor algorithms and multiple sequence alignments. J. Mol. Biol. 1995, 247: 11-15

[28] Schapire, R.E. A brief introduction to boosting. In Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence 1999, 1401-1406

[29] Songyang, Z., Shoelson, S.E., Chaudhuri, M., Gish, G., Pawson, T., Haser, W.G., King, F., Roberts, T., Ratnofsky, S., Lechleider, R.J., Neel, B.G., Birge, R.B., Fajardo, J.E., Chou, M.M. and Hanafusa, H. et al. SH2 domains recognize specific phosphopeptide sequences. Cell 1993, 72:767-778

[30] Vapnik, V. Statistical Learning Theory. John Wiley, New York, 1998

[31] Xue, Y., Zhou, F., Zhu, M., Ahmed, K., Chen, G. and Yao, X. GPS: a comprehensive www server for phosphorylation sites prediction. Nucleic Acids Research 2005, 33:W184–W187

[32] Yaffe, M.B., Leparc, G.G., Lai, J., Obata, T., Volinia, S. and Cantley, L.C. A motif-based profile scanning approach for genome-wide prediction of signaling pathways. Nat. Biotechnol 2001, 19:348-353

[33] Zanzoni, A., Ausiello, G., Via, A., Gherardini, P.F., and Helmer-Citterich, M. Phospho3D: a database of three-dimensional structures of protein phosphorylation sites. Nucl. Acids Res. 2007, 35: D229-D231

[34] Zvelebil, M.J., Barton, G.J., Taylor, W.R., and Sternberg, M.J.E. Prediction of protein secondary structure and active sites using alignment of homologous sequence. J. Mol. Biol. 1987, 194: 957-961