Predicting Palmitoylation Sites Using a Regularised Bio-basis Function Neural Network

Zheng Rong Yang

School of Engineering, Computer Science and Mathematics
University of Exeter, UK

Abstract. Palmitoylation is one of the most important post-translational modifications involving molecular signalling activities. Two simple methods have been developed very recently for predicting palmitoylation sites, but the sensitivity (the prediction accuracy of palmitoylation sites) of both methods is low ($< 65\%$). A regularised bio-basis function neural network is implemented in this paper aiming to improve the sensitivity. A set of protein sequences with experimentally determined palmitoylation sites are downloaded from NCBI for the study. The protein-oriented cross-validation strategy is used for proper model construction. The experiments show that the regularised bio-basis function neural network significantly outperforms the two existing methods as well as the support vector machine and the radial basis function neural network. Specifically the sensitivity has been significantly improved with a slightly improved specificity (the prediction accuracy of non-palmitoylation sites).

Keywords: Palmitoylation site prediction, bio-basis function, regularisation.

1 Introduction

Palmitoylation is a hydrophobic protein-modification activity where fatty acids are covalently attached to cysteine residues of membrane proteins. In biochemistry and enzymology study, it has been observed that this hydrophobic protein-modification activity uses cellular and viral membrane proteins for signal transmission [1]. It is still unknown what the molecular signals for palmitoylation are. Although palmitoylation is known to be a reversible activity with cycles of acylation and deacylation, the relevant enzymatic mechanism has not been completely known because some palmitated proteins are found without any enzyme source present. Despite of these observations, palmitoylation activity has been widely studied in various areas including most signalling pathway activities [2], [3]. For instance, Smotrys et al showed that most trafficking and protein-protein interactions as well as enzyme activities depend on the existence of palmitated proteins [4]. They also showed that palmitated proteins can enhance the membrane interactions and the reversibility of palmitoylation is an attractive mechanism for regulating protein activity and cell signalling. Li and Yang have found that most palmitoylation-deficient mutant Env proteins are

I. Măndoiu and A. Zelikovsky (Eds.): ISBRA 2007, LNBI 4463, pp. 406–417, 2007.

soluble when extracted by ice-cold TX-100 and stay at the bottom of the gradients in their study of the association between the Maurine leukaemia virus Env protein and lipid rafts [5]. Palmitoylation has also been studied in disease-related subjects. For instance, Yu and Lee have found that two cysteine residues (257 and 261) at the C-terminal of NS4B have lipid modifications (palmitoylation) in studying the polymerization of Hepatitis C virus NS4B protein [6]. They concluded that site-specific mutagenesis of these cysteine residues are important for protein-protein interactions in the formation of HCV RNA replication complex. Another example of disease-related biology study of palmitoylation activity is vacuolar events. Peng and Tang showed that palmitoylation targets Vac8p to specific membrane sub-domains for vacuole homotypic fusion at three cysteine residues (4, 5 and 7) [7].

Specificity study of post-translational modifications like phosphorylation, methylation, sumoylation and palmitoylation is a very important subject in systems biology research for understanding how proteins are responding to extracellular cues for information transmission along signalling pathways. One of the important subjects in studying post-translational modifications is to identify where the modifications are or where proteins are binding for the modifications. For this kind of study, it is generally not necessary to view whole protein sequences. Rather, one normally focuses on a small area of a binding site or a few residues around a functional site. This study is commonly termed as protein functional site prediction which involves the use of a set of peptides (short regions of protein sequences) with known functional status, i.e. functional or non-functional. In this context, a functional peptide is the one with a palmitoylation site.

The earliest work on protein functional site prediction were normally based on frequency estimate. For example, the h function [8], where the frequency of 20 amino acids at each residue is calculated from a set of functional peptides. The estimated frequencies are then stored in a computer program for prediction. The major shortcoming of this method is that they usually result in high sensitivity and low specificity. Some statistical models like hidden Markov models (HMM) [9], discriminant analysis [10] and quadratic discriminant analysis [11] have also been used for data mining protein peptides. However, a HMM model also has a high sensitivity and a low specificity [12].

Neural networks and the support vector machine [13], [14] have been applied to data mining protein peptides as well. For instance, neural networks have been used in signal peptide cleavage site prediction [15], glycoproteins linkage site prediction [16], enzyme active site prediction [17], phosphorylation site prediction [18], and water active site prediction [19]. The support vector machine has been used for the prediction of translation initiation sites [20], the prediction of phosphorylation sites [21], the prediction of T-cell receptor [22], and the prediction of protein-protein interactions [23].

In the context of predicting palmitoylation sites, Zhou et al first employed a clustering and scoring strategy to build a model to predict palmitoylation sites in early 2006 [24]. In the same group, Xue et al employed a Naive Bayes method to predict palmitoylation sites in late 2006 [25]. They have used 105 protein

sequences with 245 palmitoylation sites being experimentally determined. Based on these protein sequences, 977 non-palmitoylation peptides with various lengths were generated, each having a cysteine in the middle. Their model [25] is able to make total prediction accuracy 86.74% with the sensitivity 58.37%, the specificity 93.86% and the Matthews' correlation coefficient 0.5618. In comparison, they have used the support vector machine and the radial basis function neural network. For the former, the specificity is 94.47%, the sensitivity is 64.49% and the Matthews' correlation coefficient is 0.623. For the latter, the sensitivity is 95.09%, the specificity is 55.51% and the Matthews' correlation coefficient is 0.5664. It can be seen that all have a low sensitivity from 58.37% to 64.49%. For the Naive Bayes method, they found that the optimal window size is six. For the support vector machine, they found that the optimal window size is seven. For the radial basis function neural network, they found that the best window size is eight. In dealing with amino acids, they employed the orthogonal coding mechanism where each amino acid is coded using a 20-bit long orthogonal binary vector [26].

Although the orthogonal coding mechanism has been widely used for various protein peptide modelling tasks, it may not well code biological information in peptides. The bio-basis function neural network was therefore developed for proper coding of amino acids in 2003 [27], [28]. The bio-basis function neural network has been successfully used for Trypsin cleavage site prediction [27], HIV cleavage site prediction [28], [30], [29], disordered protein prediction [31], [32], phosphorylation site prediction [33], [12], glycoprotein O-linkage site prediction [34], Caspase cleavage site prediction [35], SARS-CoV protease cleavage site prediction [36], signal peptide prediction [37], [38], and T-cell epitope prediction [39]. In all these applications, no regularisation was applied. This means that the models may possibly overfit to the training peptides. With the regularisation theory [40], we can constrain the model parameters to trade off between bias and variance so as to improve model generalisation capability when a proper regularisation constant is determined. This has been widely studied in neural network community [41], [42].

In this study, a regularised bio-basis function neural network is implemented for improving palmitoylation site prediction sensitivity using the data downloaded from NCBI. First, the regularised bio-basis function neural network is introduced and then how data downloaded from NCBI are organised for simulation is discussed. Particularly, the protein-oriented cross-validation strategy is discussed for proper model construction. A comparison will be given showing if the regularised bio-basis function neural network can improve the sensitivity for palmitoylation site prediction.

2 Regularised Bio-basis Function Neural Network

Before discussing the regularised bio-basis function neural network, the bio-basis function neural net is briefly discussed. Given two peptides $\mathbf{s}_i$ and $\mathbf{s}_j$,

the likelihood that they are from the same ancestor through evolution is $\mathcal{L} = \prod_{k=1}^{d} p(s_{ik}, s_{jk})$. Here d is the number of residues in two peptides, $p(s_{ik}, s_{jk})$ is the probability that both s_{ik} and s_{jk} occur in $\mathbf{s}_i$ and $\mathbf{s}_j$ at the same time. Applying a logarithm operation on the likelihood function leads to

$$\rho(\mathbf{s}_i, \mathbf{s}_j) = \ln \mathcal{L} = \sum_{k=1}^{d} \mathcal{M}(s_{ik}, s_{jk}) \tag{1}$$

Here $\mathcal{M}(s_{ik}, s_{jk})$ can be found from various mutation matrices [43], [44], [45]. The bio-basis function is designed as follows

$$\phi(\mathbf{s}_i, \mathbf{s}_j) = \exp\left(\frac{\rho(\mathbf{s}_i, \mathbf{s}_j) - \rho(\mathbf{s}_j, \mathbf{s}_j)}{\rho(\mathbf{s}_j, \mathbf{s}_j)}\right) \tag{2}$$

It can be seen that $\phi(\mathbf{s}_i, \mathbf{s}_j) \in (0, 1]$. When $\mathbf{s}_i = \mathbf{s}_j$, $\phi(\mathbf{s}_i, \mathbf{s}_j) = 1$. When $\mathbf{s}_i$ is very different from $\mathbf{s}_j$, $\phi(\mathbf{s}_i, \mathbf{s}_j) \to 0$. By using the log-odds-ratio, a model using the bio-basis function for peptide classification is defined as

$$y_n = \frac{1}{1 + \exp\left(-\mathbf{w}^T \phi_n\right)} \tag{3}$$

Here $\mathbf{w} = (w_0, w_1, \cdots, w_\ell)^T$ and $\phi_n = (1, \phi_{n1}, \phi_{n2}, \cdots, \phi_{n\ell})^T$. The objective function with an added regularisation term (neg log pdf + regularisation) is then defined as

$$\mathcal{O} = -\sum_{n=1}^{\ell} \left(t_n \log y_n + (1 - t_n) \log(1 - y_n)\right) + \frac{1}{2}\lambda \mathbf{w}^T \mathbf{w} \tag{4}$$

Here λ is a regularisation constant. The update rule of the parameters is defined as

$$\Delta \mathbf{w} = -\mathbf{H}^{-1} \nabla \mathcal{O} = \left(\mathbf{\Phi}^T \mathbf{\Lambda} \mathbf{\Phi} + \lambda \mathbf{I}\right)^{-1} \left(\mathbf{\Phi}^T \mathbf{e} - \lambda \mathbf{w}\right) \tag{5}$$

Here $\mathbf{I}$ is an identity matrix, $\mathbf{\Lambda} = \mathrm{diag}\{y_n(1 - y_n)\}$ is called an entropy matrix, $\nabla \mathcal{O}$ is the first derivative of $\mathcal{O}$ with respect to $\mathbf{w}$,

$$\mathbf{H} = \nabla \nabla \mathcal{O} \tag{6}$$

is the Hessian matrix, $\mathbf{e} = (e_1, e_2, \cdots, e_\ell)^T$, $e_n = t_n - y_n$, and

$$\mathbf{\Phi} = \begin{pmatrix} 1 & \phi_{11} & \phi_{12} & \cdots & \phi_{1\ell} \\ 1 & \phi_{21} & \phi_{22} & \cdots & \phi_{2\ell} \\ 1 & \vdots & \vdots & \vdots & \vdots \\ 1 & \phi_{\ell 1} & \phi_{\ell 2} & \cdots & \phi_{\ell\ell} \end{pmatrix} \tag{7}$$

The learning procedure is designed as below

[1]: $c = 0$, $\mathbf{w}^c = \mathbf{0}$
[2]: Calculate $\mathbf{y}^c = (y_1^c, y_2^c, \cdots, y_\ell^c)^T$, $\mathbf{e}^c$ and $\mathbf{\Lambda}^c$

[3]: $\mathbf{w}^{c+1} = \mathbf{w}^c + \left(\mathbf{\Phi}^T \mathbf{\Lambda}^c \mathbf{\Phi} + \lambda \mathbf{I}\right)^{-1} \left(\mathbf{\Phi}^T \mathbf{e}^c - \lambda \mathbf{w}^c\right)$
[4]: If $\|\mathbf{w}^{c+1} - \mathbf{w}^c\| < \epsilon$, stop, otherwise $c = c + 1$, goto [2].

Here $\mathbf{w}^c$ is $\mathbf{w}$ at c^{th} learning cycle, $\mathbf{y}^c$ is $\mathbf{y}$ at c^{th} learning cycle, $\mathbf{e}^c$ is $\mathbf{e}$ at c^{th} learning cycle, $\mathbf{\Lambda}^c$ is $\mathbf{\Lambda}$ at c^{th} learning cycle and $\epsilon > 0$ is a small number functioning as a termination rule. The other termination rule is the maximum training cycle (being set 100 in this paper). In most cases, the first termination rule is satisfied.

3 Result

3.1 Data

A data set of 55 protein sequences with 90 experimentally confirmed palmitoylation sites was downloaded from NCBI. It has been found that palmitoylation activity won't happen if cysteine is not present. We can then scan these 55 protein sequences to generate peptides with cysteine in the middle (P_0). In total, there are 490 cysteine residues in these 55 protein sequences. There are on average 8.9 cysteine residues in each protein sequence and less than two of them are possible palmitoylation sites. The peptide chain with $2n+1$ residues is expressed as $\mathrm{P}_n - \cdots - \mathrm{P}_1 - \mathrm{P}_0 - \mathrm{P}_{1'} - \cdots - \mathrm{P}_{n'}$.

3.2 Cross-Validation

The next question is then how to use the data for constructing models for prediction. We don't want a model to overfit the training data in any circumstances. Cross-validation or jackknife is certainly a way for achieving this goal. However, the fundamental principle of cross-validation has been very often abused in data mining protein peptides. In using cross-validation, one important principle is that we cannot use a data set which has any information exposed to training for evaluation. If this happens, the model can be very likely over-evaluated. However, in many applications, this important issue has not been seriously addressed. Subsequences or peptides are normally generated through scanning all the available protein sequences at first. These generated peptides are then pooled together and then randomly divided for cross-validation. With this strategy, we may almost over-evaluate a model or an algorithm. The reason is very simple. Each protein sequence can be treated as a small world in which mutation (although the underline mechanism of it has not yet been completely known) happens in a specific way which may differ from other protein sequences. If we have generated peptides before cross-validation, some peptides generated from a protein sequence can be randomly picked up for training and some peptides generated from the same protein sequence can be randomly picked up for testing. This means that the pattern in testing peptides have already partially known in training! To handle this problem, we have proposed a new strategy called protein-oriented cross-validation [49], [50]. The core principle of the protein-oriented cross-validation is to divide protein sequences into k folds at first. For each sequence in each fold, a sliding window is applied to generate peptides. Cross-validation simulation is then run based on peptides in these folds.

3.3 Sequence Logos

Fig 1 and 2 show the sequence logos for palmitated and non-palmitated peptides. The logos were produced using the WebLogo[1] [47]. Note that the middle cysteine residue is removed. This means that we are working for the peptides in the following format $P_n - \cdots - P_1 - P_{1'} - \cdots - P_{n'}$. From Fig 1 and 2, it can be seen that two classes of peptides show some difference in amino acid distributions.

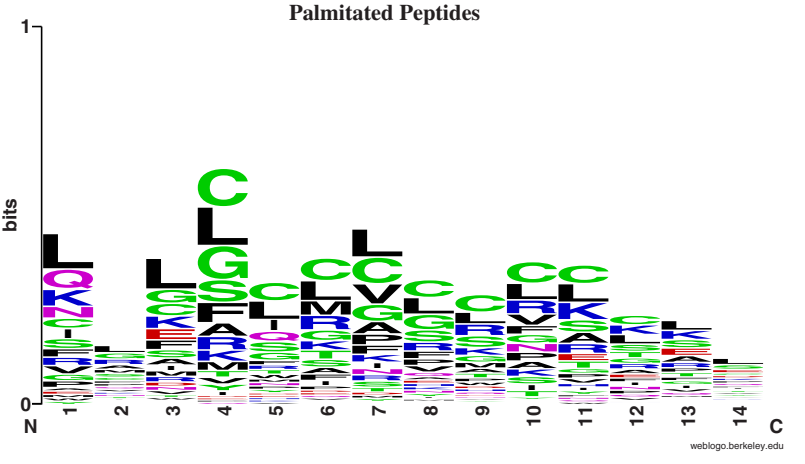


Fig. 1. Sequence logos generated for palmitated peptides

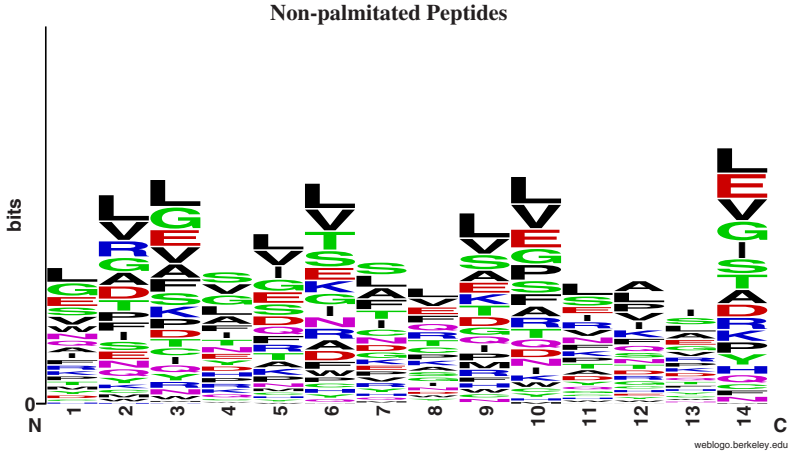


Fig. 2. Sequence logos generated for non-palmitated peptides

[1] http://weblogo.berkeley.edu/logo.cgi

3.4 Model Evaluation

Two criteria are used for model evaluation. They are the Matthews' correlation coefficient [46] and the receiver operating characteristic (ROC) curve [48]. Let TN, TP, FN, FP denote true negative (correctly identified non-palmitated peptides), true positive (correctly identified palmitated peptides), false negative (palmitated peptides identified as non-palmitated ones) and false positive (non-palmitated peptides identified as palmitated ones), respectively. The Matthews' correlation coefficient (MCC) is

$$MCC = \frac{TN \times TP - FN \times FP}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{8}$$

The Matthews' correlation coefficient measures how the predictions correlate with the real target values. If the coefficient is positive, the predictions are positively correlated with the target values. If the Matthews' correlation coefficient is zero, the prediction is completely random. For the ROC analysis, we use the area under a ROC curve (AUR) for the testing set as it is a quantitative measurement of the robustness of a built model.

3.5 Result

For simulation, we have changed the λ (see Eq. 4) value from the range (0.001, 0.002, 0.004, 0.006, 0.008, 0.01, 0.02, 0.04, 0.06, 0.08). The simulation result using 10 λ values for 6 window sizes (5, 7, 9, 11, 13, and 15) are shown in Fig 3

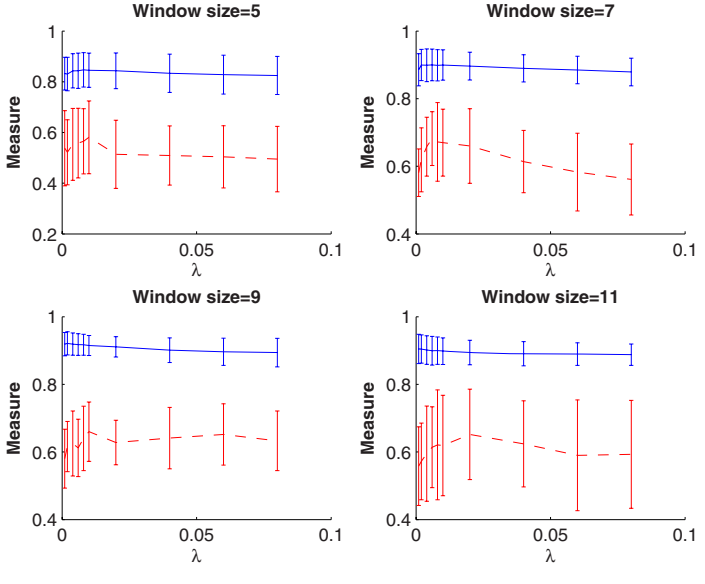


Fig. 3. Performance for window sizes 5, 7, 9, and 11

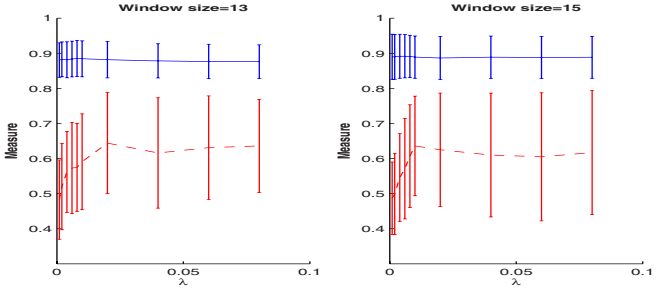
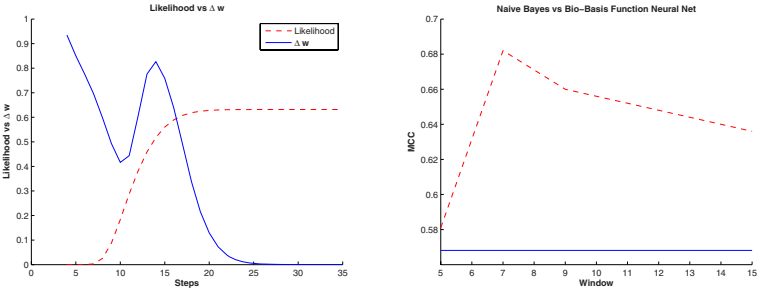


Fig. 4. Performance for window sizes 13 and 15



(a) (b)

Fig. 5. (a) Learning performance using the regularised bio-basis function neural network. (b) MCC comparison between Naive Bayes and Bio-Basis Function Neural Net.

and Fig 4, where the dashed lines are for MCC and the solid lines are for AUR. It can be seen for almost all cases, small λ values produce better models.

Fig 5 (a) shows the learning performance using the regularised bio-basis function neural network. It can be seen that the likelihood $\prod_{n=1}^{\ell} y_n^{t_n}(1-y_n)^{1-t_n}$ is consistently increasing and the change of weights $\|\Delta \mathbf{w}\|$ is decreasing.

3.6 Comparison

Table 1 shows a comparison between different algorithms. It can be seen that the regularised bio-basis function neural net significantly outperforms the Naive Bayes, the support vector machine and the radial-basis function neural net. The best bio-basis function neural network uses 7-mer (in fact 6-mer after removing P_0) peptides and $\lambda = 0.006$. The best bio-basis function neural net and the Naive Bayes show a difference 0.12 in MCC. The best bio-basis function neural net and the Naive Bayes show a difference 0.09% in specificity. However, the best bio-basis function neural net and the Naive Bayes show a difference 16.04% in sensitivity accounting for 27.5% increase! Fig 5 (b) shows a MCC comparison between the Naive Bayes and the bio-basis function neural network, where the dash line represents the bio-basis function neural network using various window

sizes while the straight solid line represents the performance using the Naive Bayes method. It can be seen that the latter much outperforms the former.

It has been mentioned above that the best window sizes are six using the Naive Bayes method, seven using the support vector machine, eight using the radial-basis function neural network. The regularised bio-basis function neural network selects the best window size as seven as the support vector machine which produced the best sensitivity compared with the rest conventional methods. It should be noted that both the support vector machine and the regularised bio-basis function neural network use the regularisation theory to improve generalisation capability. This may be the reason that these two are close in the prediction sensitivity and the window size.

Table 1. The comparison between different algorithms. Naive Bayes, Support vector machine and Radial-basis function neural net have no report of AUR measures. The number within brackets mean the window size used for modelling.

Algorithms	λ	Specificity	Sensitivity	MCC	AUR
Naive Bayes		93.86%	58.37%	0.562	n.a.
Support vector machine		94.47%	64.44%	0.623	n.a.
Radial-basis function neural net		95.09%	55.51%	0.537	n.a.
Bio-basis function neural net (5)	0.01	89.78%	70.10%	0.581	0.845
Bio-basis function neural net (7)	0.006	95.95%	74.41%	0.682	0.899
Bio-basis function neural net (9)	0.01	93.70%	72.53%	0.660	0.920
Bio-basis function neural net (11)	0.02	95.41%	63.39%	0.652	0.894
Bio-basis function neural net (13)	0.02	93.66%	68.67%	0.644	0.882
Bio-basis function neural net (15)	0.01	89.88%	77.24%	0.636	0.889

It should be noted that the models presented by Zhou et al [24] and Xue et al [25] used 245 palmitated peptides and 977 non-palmitated peptides compared with 90 palmitated peptides and 400 non-palmitated peptides. The author is contacting Zhou et al and Xue et al at the moment for requesting their data. It is expected that the regularised bio-basis function neural network will even perform better after their data arrive.

4 Conclusion

This paper has implemented a regularised bio-basis function neural network for predicting palmitoylation sites in proteins. Through comparison, it has been found that the new method presented in this paper significantly outperforms the traditional methods, namely the Naive Bayes method, the support vector machine and the radial-basis function neural network. We are currently investigating how to use the regularised bio-basis function neural network to produce sparse models for better interpretation to the trained models. In using the information provided by the Hessian matrix, we can evaluate the importance of each bio-basis using the statistic as $Z_n = \frac{w_n}{\sqrt{H_{nn}^{-1}}}$ $(\forall n \in [0, \ell])$. If $Z_n < \vartheta$ $(\vartheta > 0$

is a small number functioning as a threshold), w_n can be zeroed or the n^{th} bio-basis can be removed. A detailed research is undergoing for investigating how to determine ϑ and how to interpret the left bio-bases in biology.

References

1. Veit, M., Schmidt, M.F.G.: Palmitoylation of viral and cellular proteins. In: Influenza Viruses. Facts and Perspectives, Schmidt, Michael F.G.(Hrsg.) Berlin: Grosse-Verlag ISBN: 3-9810221-3-0 (2006)
2. Navarro-Lerida, I., Alvarez-Barrientos, A., Rodrguez-Crespo, I.: N-terminal palmitoylation within the appropriate amino acid environment conveys on NOS2 the ability to progress along the intracellular sorting pathways. Journal of Cell Science **119** (2006) 1558–1596
3. Kurayoshi, M., Yamamoto, H., Izumi, S., Kikuchi, A.: Post-translational palmitoylation and glycosylation of Wnt-5a are necessary for its signaling
4. Smotrys, J.E., Linder, M.E.: Palmitoylation of intracellular signaling proteins: regulation and function. Annu. Rev. Biochem. **73** (2004) 559–587
5. Li, M., Yang, C., Tong, C., Weidmann, A., Compans, R.W.: Palmitoylation of the murine leukemia virus envelope protein is critical for lipid raft association and surface expression. J Virol. **76** (2002) 11845–11852
6. Yu, G., Lee, K., Gao, L., Lai, M.M.C.: Palmitoylation and Polymerization of Hepatitis C Virus NS4B Protein. Journal of Virology **80** (2006) 6013–6023
7. Peng, Y., Tang, F., Weisman, L.S.: Palmitoylation plays a role in targeting Vac8p to specific membrane subdomains. Traffic **7** (2006) 1378
8. Poorman, R.A., Tomasselli, A.G., Heinrikson, R.L., Kezdy, F.J.: A cumulative specificity model for protease from human immunodeficiency virus types 1 and 2, inferred from statistical analysis of an extended substrate data base. J. Biol. Chem **22** (1991) 14554–14561
9. Rabiner, L.R.: A tutorial on hidden Markov models and selected applications in speech recognition. Proc. IEEE **77** (1989) 257–286
10. Nakata, K., Maizel, J.V.: Prediction of operator-binding protein by discriminant analysis. Gene Anal Tech **6** (1989) 111–119
11. Chen, C.P., Rost, B.: State-of-the-art in membrane protein prediction. Applied Bioinformatics **1** (2002) 21–35
12. Senawongse, P., Dalby, A., Yang, Z.R.: Predicting the phosphorylation sites using hidden Markov models and Machine Learning methods. Journal of Chemical Information and Computer Science **45** (2005) 1147–1152
13. Vapnik, V.: The Nature of Statistical Learning Theory.Springer-Verlag, New York (1995)
14. Scholkopf, B.: The kernel trick for distances, Technical Report. Microsoft Research May (2000)
15. Nielsen, M., Lundegaard, C., Worning, P., Lauemoller, S.L., Lamberth, K., Buss, S., et al. Reliable prediction of T-cell epitopes using neural networks with novel sequence representations. Protein Science **12** (2003) 1007–1017
16. Hansen, J.E., Lund, O., Engelbrecht, J., Bohr, H., Nielsen, J.O.: Prediction of O-glycosylation of mammalian proteins: specificity patterns of UDP-GalNAc: polypeptide N-acetylgalactosaminyltransferase. Biochem J. **30** (1995) 801–813
17. Gutteridge, A., Bartlett, G.J., Thornton, J.M.: Using a neural network and spatial clustering to predict the location of active sites in enzymes. Journal of Molecular Biology **330** (2003) 719–734

18. Blom, N., Gammeltoft, S., Brunak, S.: Sequence and structure based prediction of eukaryotic protein phosphorylation sites. J. Mol. Biol **24** (1999) 1351–1362
19. Ehrlich, L., Reczko, M., Bohr, H., Wade, R.C.: Prediction of protein hydration sites from sequence by modular neural networks. Protein Eng **11** (1998) 11–19
20. Zien, A., Ratsch, G., Mika, S., Scholkopf, B., Lengauer, T. and Muller, K.R.: Engineering support vector machine kernels that recognize translation initiation sites. Bioinformatics **16** (2000) 799–807
21. Kim, J.H., Lee, J., Oh, B., Kimm, K., Koh, I.: Prediction of phosphorylation sites using SVMs. Bioinformatics **20** (2006) 3179–3184
22. Zhao, Y., Pinilla, C., Valmori, D., Martin, R., Simon, R.: Application of support vector machines for T-cell epitopes prediction. Bioinformatics **19** (2003) 1978–1984
23. Koike, A., Takagi, T.: Prediction of protein-protein interaction sites using support vector machines. Protein Eng Des Sel **17** (2004) 165–173
24. Zhou, F., Xue, Y., Yao, X., Xu, Y.: CSS-Palm: palmitoylation site prediction with a clustering and scoring strategy (CSS). Bioinformatics **22** (2006) 894–896
25. Xue, Y., Chen, H., Jin, C., Sun, Z., Yao, X.: NBA-Palm: prediction of palmitoylation site implemented in Nave Bayes algorithm. BMC Bioinformatics **7** (2006) 1–10
26. Qian, N., Sejnowski, T.: Predicting the secondary structure of globular proteins using neural network models. Proceeding of Int J. Conf. On Neural Networks, (1998) 865–884
27. Thomson, R., Hodgman, T., Yang, Z.R., Doyle, A.: Characterising proteolytic cleavage site activity using bio-basis function neural networks. Bioinformatics **19** (2003) 1741–1747
28. Yang, Z.R., Thomson, R.: Bio-basis function neural network for prediction of protease cleavage sites in proteins. IEEE Trans. on Neural Networks **16** (2005) 263–274
29. You, L., Garwicz, D., Rognvaldsson, T.: Comprehensive bioinformatic analysis of the specificity of human immunodeficiency virus type 1 protease. Journal of Virology **79** (2005) 12477–12486
30. Yang, Z.R., Berry, E.: A novel neural learning algorithm for protease cleavage site prediction. Journal of Bioinformatics and Computational Biology **2** (2004) 511–531
31. Thomson, R., Esnouf, R.: Predict disordered proteins using bio-basis function neural networks. Lecture Notes in Computer Science **3177** (2004) 19–27
32. Yang, Z.R., Thomson, R., McNeil, P., Esnouf, R.: RONN: use of the bio-basis function neural network technique for the detection of natively disordered regions in proteins. Bioinformatics **21** (2005) 3369–3376
33. Berry, E., Dalby, A., Yang, Z.R.: Reduced bio-basis function neural networks in prediction of phosphorylation sites, a comparative study. Computational Biology and Chemistry **28** (2004) 75–85
34. Yang, Z.R., Chou, K.C.: Predicting the O-linkage sites in glycoproteins using bio-basis function neural networks. Bioinformatics **20** (2004) 903–908
35. Yang, Z.R.: Prediction of caspase cleavage sites using Bayesian bio-basis function neural networks. Bioinformatics **21** (2005) 1831–1837
36. Yang, Z.R.: Mining SARS-CoV protease cleavage data using decision trees, a novel method for decisive template searching. Bioinformatics **21** (2005) 2644–2650
37. Sidhu, A., Yang, Z.R.: Prediction of signal peptides using bio-basis function neural networks and decision trees. Applied Bioinformatics **5** (2006) 13–19
38. Yang, Z.R.: Orthogonal kernel machine in prediction of functional sites in preteins. IEEE Trans on Systems, Man and Cybernetics **35** (2005) 100–106

39. Yang, Z.R., Johnathan, F.: Predict T-cell epitopes using bio-support vector machines. Journal of Chemical Information and Computer Sciences **45** (2005) 1142–1148
40. Neumaier, A.: Solving ill-conditioned and singular linear systems: A tutorial on regularization, SIAM Review **40** (1998) 636–666
41. Girosi, F., Jones, M., Poggio, T.: Regularization Theory and Neural Networks Architectures Neural Computation **7** (1995) 219–269
42. Bishop, C.: Neural Networks for Pattern Recognition, Oxford Press, 1995
43. Dayhoff, M.O., Schwartz, R.M., Orcutt, B.C.: A model of evolutionary change in proteins. matrices for detecting distant relationships. Atlas of protein sequence and structure **5** (1978) 345–358
44. Henikoff, S., Henikoff, J.G.: Amino acid Substitution matrices from protein blocks. Proc. Natl. Acad. Sci. **89** (1992) 10915–10919
45. Johnson, M.S., Overington, J.P.: A structural basis for sequence comparisons-an evaluation of scoring methodologies. Journal Molecular Biology **233** (1993) 716–738
46. Matthews, B.W.: Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta **405** (1975) 442–451
47. Schneider, T.D., Stephens, R.M.: Sequence Logos: A new way to display consensus sequences. Nucleic Acids Res. **18** (1990) 6097–6100
48. Metz, C.E.: Basic principles of ROC analysis. Seminars in Nuclear Medicine **8** (1978) 283–298
49. Yang, Z.R.: Predicting Hepatitis C virus protease cleavage sites using generalised linear indicator regression models. IEEE Trans on Biomedical Engineering. **53** (2006) 2119–2123
50. Yang, Z.R.: A probabilistic peptide machine for predicting Hepatitis C virus protease cleavage sites. IEEE Trans on Information Technology in Biomedicine (in press)