Bio-kernel Self-organizing Map for HIV Drug Resistance Classification

Zheng Rong Yang and Natasha Young

Department of Computer Science, University of Exeter,
Exeter EX4 4QF, UK
z.r.yang@ex.ac.uk
http://www.dcs.ex.ac.uk/~zryang

Abstract. Kernel self-organizing map has been recently studied by Fyfe and his colleagues [1]. This paper investigates the use of a novel bio-kernel function for the kernel self-organizing map. For verification, the application of the proposed new kernel self-organizing map to HIV drug resistance classification using mutation patterns in protease sequences is presented. The original self-organizing map together with the distributed encoding method was compared. It has been found that the use of the kernel self-organizing map with the novel bio-kernel function leads to better classification and faster convergence rate…

1 Introduction

In analysing molecular sequences, we need to select a proper feature extraction which can convert the non-numerical attributes in sequences to numerical features prior to using a machine learning algorithm. Suppose we denote by $\mathbf{x}$ a sequence and $\phi(\mathbf{x})$ a feature extraction function, the mapping using a feature extraction function is $F(F: S \to \phi) \in R^d$. Finding an appropriate feature extraction approach is a non-trivial task.

It is known that each protein sequence is an ordered list of 20 amino acids while a DNA sequence is an ordered list of four nucleic acids. Both amino acids and nucleic acids are non-numerical attributes. In order to analyze molecular sequences, these non-numerical attributes must be converted to numerical attributes through a feature extraction process for using a machine learning algorithm. The distributed encoding method [2] was proposed in 1988 for extracting features for molecular sequences. The principle is to find orthogonal binary vectors to represent amino (nucleic) acids. With this method, amino acid Alanine is represented by 0000000000 0000000001 while Cystine 0000000000 0000000010, etc. With the introduction of this feature extraction method, the application of machine learning algorithms to bioinformatics has been very successful. For instance, this method has been applied to the prediction of protease cleavage sites [3], signal peptide cleavage sites [4], linkage sites in glycoproteins [5], enzyme active sites [6], phosphorylation sites [7] and water active sites [8].

However, as indicated in the earlier work [9], [10], [11] such a method has its inherent limit in two aspects. First, the dimension of an input space has been enlarged 20 times weakening the significance of a set of training data. Second, the biological

L. Wang, K. Chen, and Y.S. Ong (Eds.): ICNC 2005, LNCS 3610, pp. 179 – 186, 2005.

content in a molecule sequence may not be efficiently coded. This is because the similarity between any pair of different amino (nucleic) acids varies while the distance between such encoded orthogonal vectors of two different amino (nucleic) acids is fixed.

The second method for extracting features from protein sequences is to calculate the frequency. It has been used for the prediction of membrane protein types [12], the prediction of protein structural classes [13], subcellular location prediction [14] and the prediction of secondary structures [15]. However, the method ignores the coupling effects among the neighbouring residues in sequences leading to potential bias in modelling. Therefore, di-peptides method was proposed where the frequency of each pair of amino acids occurred as neighbouring residues is counted and is regarded as a feature. Dipeptides, gapped (up to two gaps) transitions and the occurrence of some motifs as additive numerical attributes were used for the prediction of subcellular locations [16] and gene identification [17]. Descriptors were also used, for instance, to predict multi-class protein folds [18], to classify proteins [19] and to recognise rRNA-, RNA-, and DNA-binding proteins [20], [21]. Taking into account the high order interaction among the residues, multi-peptides can also be used. It can be seen that there are 400 di-peptides, 8,000 tri-peptides and 16,000 tetra-peptides. Such a feature space can be therefore computational impractical for modelling.

The third class of methods is using profile measurement. A profile of a sequence can be generated by subjecting it to a homology alignment method or Hidden Markov Models (HMMs) [22], [23], [24], [25].

It can be seen that either finding an appropriate approach to define $\phi(\mathbf{x})$ is difficult or the defined approach may lead to a very large dimension, i.e., $d \to \infty$. If an approach which can quantify the distance or similarity between two molecular sequences is available, an alternative learning method can be proposed to avoid the difficulty in searching for a proper and efficient feature extraction method. This means that we can define a reference system to quantify the distance among the molecular sequences. With such a reference system, all the sequences are quantitatively featured by measuring the distance or similarity with the reference sequences.

One of the important issues in using machine learning algorithms for analysing molecular sequences is investigating sequence distribution or visualising sequence space. Self-organizing map [26] has been one of the most important machine learning algorithms for this purpose. For instance, SOM has been employed to identify motifs and families in the context of unsupervised learning [27], [28], [29], [30], [31]. SOM has also been used for partitioning gene data [32]. In these applications, feature extraction methods like the distributed encoding method were used.

In order to enable SOM to deal with complicated applications where feature extraction is difficult, kernel method has been introduced recently by Fyfe and his colleagues [1]. Kernel methods were firstly used in cluster analysis for K-means algorithms [33], where the Euclidean distance between an input vector $\mathbf{x}$ and a mean vector $\mathbf{m}$ is minimized in a feature space spanned by kernels. In the kernel feature space, both $\mathbf{x}$ and $\mathbf{m}$ were the expansion on the training data. Fyfe and his colleagues developed so-called kernel self-organizing maps [34], [35]. This paper aims to introduce a bio-kernel function for kernel SOM. The method is verified on HIV drug resistance classification. A stochastic learning process is used with a regularization term.

2 Methods

A training data set $\mathsf{D} = \{\mathbf{s}_n\}_{n=1}^{\hbar}$, where $\mathbf{s}_n \in \mathsf{S}^D$ (S is a set of possible values and $|\mathsf{S}|$ can be either definite or indefinite) and a mapping function which can map a sequence to a numerical feature vector is defined as $\mathsf{F}(\phi : \mathsf{S} \to \mathsf{F}) \in \mathrm{R}^d$, $\mathbf{x}_n = \boldsymbol{\varphi}(\mathbf{s}_n)$. In most situations, $\mathbf{x}_n = \boldsymbol{\varphi}(\mathbf{s}_n) = (\phi_1(\mathbf{s}_n), \phi_2(\mathbf{s}_n), , \phi_d(\mathbf{s}_n))^{\mathrm{T}}$ is unknown and possibly, $d \to \infty$. This then causes the difficulty in modelling. In using self-organizing map for unsupervised learning of protein sequences, the error function in the feature space F can be defined as $\mathsf{L} = |\mathbf{x}_n - \mathbf{w}_m|^2$, where $\mathbf{w}_m \in \mathrm{R}^d$ is the weight vector connecting the *m*th output neuron. Suppose $\mathbf{w}_m$ can be expanded on the training sequences ($\mathbf{w}_m = \Phi\boldsymbol{\alpha}_m$). Note that $\boldsymbol{\alpha}_m \in \mathrm{R}^{\hbar}$ is an expansion vector and $\Phi = \{\phi_i(\mathbf{s}_j)\}_{1 \le i \le d, 1 \le j \le \hbar}$. The error function can re-written as $\mathsf{L} = \mathsf{K}_{nn} - 2\mathbf{k}_n\boldsymbol{\alpha}_m + \boldsymbol{\alpha}_m^{\mathrm{T}}\mathbf{K}\boldsymbol{\alpha}_m$. Note that $\mathsf{K}_{ij} = \mathsf{K}(\mathbf{s}_i, \mathbf{s}_j)$ is the kernel, $\mathbf{k}_n = (\mathsf{K}_{n1}, \mathsf{K}_{n2}, \hbar, \mathsf{K}_{n\hbar})$ is a row kernel vector and $\mathbf{K} = \{\mathsf{K}_{ij}\}_{1 \le i,j \le \hbar}$ a kernel matrix. The error function can be as follows if we use L_2 norm regarded as a regularization term

$$\mathsf{L} = \frac{1}{2}(\mathsf{K}_{nn} - 2\mathbf{k}_n\boldsymbol{\alpha}_m + \boldsymbol{\alpha}_m^{\mathrm{T}}\mathbf{K}\boldsymbol{\alpha}_m + \lambda\boldsymbol{\alpha}_m^{\mathrm{T}}\boldsymbol{\alpha}_m),$$

where λ is the regularization factor. The update rule is then defined as $\Delta\boldsymbol{\alpha}_m = \eta(t)(\mathbf{k}_n - (\mathbf{K} + \lambda\mathbf{I})\boldsymbol{\alpha}_m)$. In designing the bio-kernel machine, a key issue is the design of an appropriate kernel function for analysing protein or DNA sequences. Similar as in [9], [10], [11], we use the bio-basis function as the bio-kernel function

$$\mathsf{K}(\mathbf{x}, \mathbf{b}_i) = \exp\left(\frac{\mathsf{M}(\mathbf{x}, \mathbf{b}_i) - \mathsf{M}(\mathbf{b}_i, \mathbf{b}_i)}{\mathsf{M}(\mathbf{b}_i, \mathbf{b}_i)}\right)$$

where $\mathbf{x}$ is a training sequence and $\mathbf{b}_i$ is a basis sequence, both have *D* residues. Note that $\mathsf{M}(\mathbf{x}, \mathbf{b}_i) = \sum_{d=1}^{D} \mathsf{M}(x_d, b_{id})$ with x_d and b_{id} and the *d*th residue in sequences. The value of $\mathsf{M}(x_d, b_{id})$ can be found in a mutation matrix [36], [37]. The bio-basis function has been successfully used for the prediction of Trypsin cleavage sites [8], HIV cleavage sites [9], signal peptide cleavage site prediction [10], Hepatitis C virus protease cleavage sites [38], disordered protein prediction [39], [40], phosphorylation site prediction [41], the prediction of the O-linkage sites in glycoproteins [42], the prediction of Caspase cleavage sites [43], the prediction of SARS-CoV protease cleavage sites [44] and the prediction of signal peptides [45].

3 Results

Drug resistance modeling is a wide phenomenon and drug resistance modeling is a very important issue in medicine. In computer aided drug design, it is desired to study

how the genomic information is related with therapy effect [46]. To predict if HIV drug may fail in therapy using the information contained in viral protease sequences is regarded as genotype-phenotype correlation. In order to discover such relationship, many researchers have done a lot of work in this area. For instance, the original self-organizing map was used on two types of data, i.e., structural information and sequence information [46]. In using sequence information, frequency features were used as the inputs to SOM. The prediction accuracy was between 68% and 85%. Instead of neural networks, statistical methods and decision trees were also used [47], [48], [49].

Data (46 mutation patterns) were obtained from [50]. Based on this data set, bio-kernel SOM was running using different value for the regularization factor. The original SOM was also used for comparison. Both SOMs used the same structure (36 output neurons) and the same learning parameters, i.e. the initial learning rate ($\eta_{\hbar} = 0.01$). Both algorithms were terminated when the mean square error was less than 0.001 or 1000 learning iterations.

Fig. 1 shows the error curves for two SOMs. It can be seen that the bio-kernel SOM (bkSOM) converged much faster with very small errors.

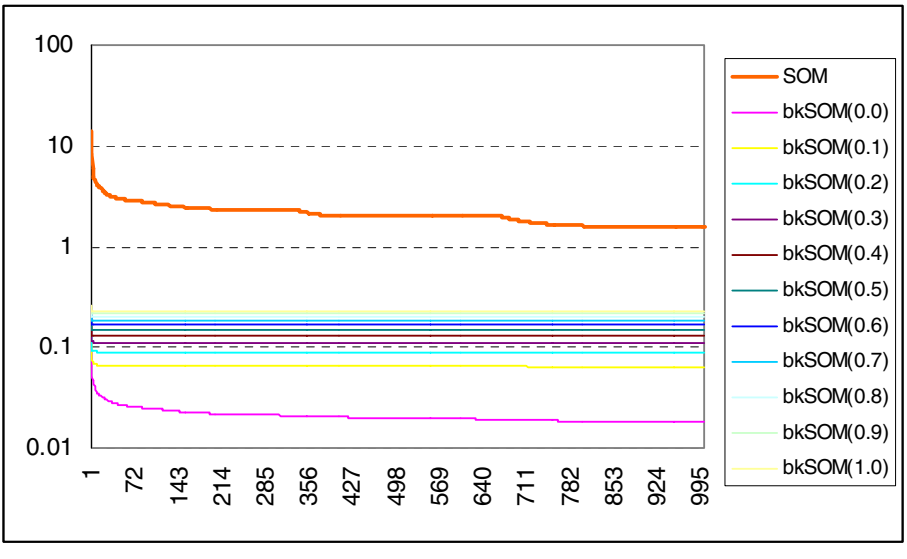

Fig. 1. The error curves for two SOMs. The horizontal axis is the learning iterations and the vertical one (logarithm scale) the errors. The numbers within the brackets of bkSOM mean the regularization factor values.

Fig. 2 shows a map of bkSOM, where "n.a." means that there is no patterns mapped onto the corresponding output neuron, "5:5" means that all the five patterns mapped onto the corresponding neuron are corresponding to the mutation patterns which are resistant to the drug and "0:9" means that all the nine patterns mapped onto the corresponding neuron are corresponding to the mutation patterns which are not resistant to the drug.

5:5	n.a.	5:5	1:1	n.a.	0:9
n.a.	1:1	n.a.	n.a.	n.a.	0:1
n.a.	n.a.	n.a.	0:1	0:3	n.a.
2:2	n.a.	n.a.	0:1	n.a.	n.a.
n.a.	0:1	0:1	1:1	n.a.	n.a.
2:2	0:1	0:2	n.a.	0:7	0:2

Fig. 2. The feature map of bkSOM.

Table 1 shows the comparison in terms of the classification accuracy, where "NR" means non-resistance and "R" resistance. It can be seen that bkSOM performed better than SOM in terms of classification accuracy. The non-resistance prediction power indicates the likelihood that a predicted non-resistance pattern is a true non-resistance pattern. The resistance prediction power therefore indicates the likelihood that a predicted resistance pattern is a true non-resistance pattern. For instance, the non-resistance prediction power using SOM is 90%. It means that for every 100 predicted non-resistance patterns, 10 would be actually resistance patterns.

Table 1. The classification accuracy of two SOMs

SOM				bkSOM			
	NR	R	Precision		NR	R	Precision
NR	28	0	100%	NR	28	0	100%
R	3	15	83%	R	0	18	100%
Power	90%	100%	93%	Power	100%	100%	100%

4 Summary

This paper has presented a novel method referred to as bio-kernel self-organizing map (bkSOM) for embedding the bio-kernel function into the kernel self-organizing map for the purpose of modeling protein sequences. The basic principle of the method is using the "kernel trick" to avoid tedious feature extraction work for protein sequences, which has been proven a non-trivial task. The computational simulation on the HIV drug resistance classification task has shown that bkSOM outperformed SOM in two aspects, convergence rate and classification accuracy.

References

1. Corchado, E., Fyfe, C. Relevance and kernel self-organising maps. International Conference on Artificial Neural Networks, (2003)
2. Qian, N., Sejnowski, T.J.: Predicting the secondary structure of globular proteins using neural network models. J. Mol. Biol., 202 (1988) 865-884
3. Thompson, T.B., Chou, K.C., Zhang, C.: Neural network prediction of the HIV-1 protease cleavage sites. Journal of Theoretical Biology, 177 (1995) 369-379.

4. Nielsen, M., Lundegaard, C., Worning, P., Lauemoller, S.L., Lamberth, K., Buss, S., Brukak, S., Lund, O.: Reliable prediction of T-cell epitopes using neural networks with novel sequence representations. Protein Science, 12 (2003) 1007- 1017
5. 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) 801-13
6. 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
7. Blom, N., Gammeltoft, S., Brunak, S.: Sequence and structure based prediction of eukaryotic protein phosphorylation sites. J. Mol. Biol. 294 (1999) 1351-1362
8. 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
9. Thomson, R., Hodgman, T. C., Yang, Z. R., Doyle, A. K.: Characterising proteolytic cleavage site activity using bio-basis function neural networks. Bioinformatics, 19 (2003) 1741-1747
10. Yang, Z.R., Thomson, R.: A novel neural network method in mining molecular sequence data. IEEE Trans. on Neural Networks, 16 (2005) 263- 274
11. Yang, Z.R.: Orthogonal kernel machine in prediction of functional sites in preteins. IEEE Trans on Systems, Man and Cybernetics, 35 (2005) 100-106
12. Cai, Y.D., Ricardo, P.W., Jen, C.H., Chou, K.C.:Application of SVMs to predict membrane protein types. Journal of Theoretical Biology, 226 (2004) 373-376
13. Cai, Y.D., Lin, X.J., Xu, X.B., Chou, K.C.:Prediction of protein structural classes by support vector machines. Computers & Chemistry, 26 (2002) 293-296
14. Hua, S., Sun, Z.: Support vector machine approach for protein subcellular localization prediction. Bioinformatics, 17 (2001) 721-728
15. Chu, F., Jin, G., Wang, L.: Cancer diagnosis and protein secondary structure prediction using support vector machines. in Wang, L. (ed) Support Vector Machines, Theory and Applications, Springer-Verlag (2004)
16. Park, K., Kanehisa, M.: Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs. Bioinformatics, 19 (2003) 1656-1663
17. Carter, R.J., Dubchak, I., Holbrook, S.R.: A computational approach to identify genes for functional RNAs in genomic sequences. Nucleic Acids Res., 29 (2001) 3928-3938
18. Ding, C.H.Q, Dubchak, I.: Multi-class protein fold recognition using support vector machines and neural networks. Bioinformatics, 17 (2001) 349-358
19. Cai, C.Z., Wang, W.L., Sun, L.Z., Chen, Y.Z.: Protein function classification via support vector machine approach. Mathematical Biosciences, 185 (2003) 111-122
20. Cai, Y.D., Lin, S.L.: Support vector machines for predicting rRNA-, RNA-, and DNA-binding proteins from amino acid sequence. Biochimica et Biophysica Acta (BBA) - Proteins & Proteomics, 1648 (2003) 127-133
21. Lin, K., Kuang, Y., Joseph, J.S., Kolatkar, P.R.: Conserved codon composition of ribosomal protein coding genes in Escherichia coli, Mycobacterium tuberculosis and Saccharomyces cerevisiae: lessons from supervised machine learning in functional genomics. Nucleic Acids Res., 30 (2002) 2599-2607
22. Jaakkola, T., Diekhans, M., Haussler, D.: Using the Fisher kernel method to detect remote protein homologies. Proceedings of the 7th International Conference on Intelligent Systems for Molecular Biology, (1999) 149-158
23. Jaakkola, T., Diekhans, M., Haussler, D.: A Discriminative Framework for Detecting Remote Protein Homologies. Journal of Computational Biology, 7 (2000) 95–114
24. Karchin, R., Karplus, K., Haussler, D.: Classifying G-protein coupled receptors with support vector machines. Bioinformatics, 18 (2002) 147-159

25. Guermeur, Y., Pollastri, G., Elisseeff, A., Zelus, D., Paugam-Moisy, H., Baldi, P.: Combining protein secondary structure prediction models with ensemble methods of optimal complexity. Neurocomputing, 56 (2004) 305-327
26. Kohonen, T.: Self organization and associative Memory, 3rd Ed (1989) Springer, Berling.
27. Arrigo, P., Giuliano, F., Scalia, F., Rapallo, A., Damiani, G.: Identification of a new motif on nucleic acid sequence data using Kohonen's self-organising map. CABIOS, 7 (1991) 353-357
28. Bengio, Y., Pouliot, Y.: Efficient recognition of immunoglobulin domains from amino acid sequences using a neural network. CABIOS, 6 (1990) 319-324
29. Ferran, E.A., Ferrara, P.: Topological maps of protein sequences. Biological Cybernetics, 65 (1991) 451-458
30. Wang, H. C., Dopazo, J., Carazo, J.M.: Self-organising tree growing network for classifying amino acids. Bioinformatics, 14 (1998) 376-377
31. Ferran, E. A. and Pflugfelder, B. A hybrid method to cluster protein sequences based on statistics and artificial neural networks. CABIOS 1993, 9, 671-680
32. Tamayo, P.; Slonim, D.; Mesirov, J.; Zhu, Q.; Kitareewan, S.; Dmitrovsky, E.; Lander, E. S. and Golub, T. R. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. PNAS 1999, 96, 2907-2912.
33. Scholkopf, B.: The kernel trick for distances, Technical Report. Microsoft Research, May (2000)
34. MacDonald, D., Koetsier, J., Corchado, E., Fyfe, C.: A kernel method for classification LNCS, 2972 (2003)
35. Fyfe, C., MacDonald, D.: Epsilon-insensitive Hebbian learning. Neuralcomputing, 47 (2002) 35-57
36. 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
37. Johnson, M.S., Overington, J.P.: A structural basis for sequence comparisons-an evaluation of scoring methodologies. J. Molec. Biol., 233 (1993) 716-738
38. Yang, Z. R., Berry, E.: Reduced bio-basis function neural networks for protease cleavage site prediction. Journal of Computational Biology and Bioinformatics, 2 (2004) 511-531
39. Thomson, R., Esnouf, R.: Predict disordered proteins using bio-basis function neural networks. Lecture Notes in Computer Science, 3177 (2004) 19-27
40. Yang Z.R., Thomson R., Esnouf R.: RONN: use of the bio-basis function neural network technique for the detection of natively disordered regions in proteins. Bioinformatics, (accepted)
41. Berry E., Dalby A. and Yang Z.R.: Reduced bio basis function neural network for identification of protein phosphorylation sites: Comparison with pattern recognition algorithms. Computational Biology and Chemistry, 28 (2004) 75-85
42. Yang, Z.R., Chou, K.C.: Bio-basis function neural networks for the prediction of the O-linkage sites in glyco-proteins. Bioinformatics, 20 (2004) 903-908
43. Yang, Z.R.: Prediction of Caspase Cleavage Sites Using Bayesian Bio-Basis Function Neural Networks. Bioinformatics (in press)
44. Yang, Z.R.: Mining SARS-CoV protease cleavage data using decision trees, a novel method for decisive template searching. Bioinformatics (accepted)
45. Sidhu, A., Yang, Z.R.: Predict signal peptides using bio-basis function neural networks. Applied Bioinformatics, (accepted)
46. Draghici, S., Potter, R.B.: Predicting HIV drug resistance with neural networks. Bioinformatics, 19 (2003) 98-107
47. Beerenwinkel, N., Daumer, M., Oette, M., Korn, K., Hoffmann, D., Kaiser, R., Lengauer, T., Selbig, J., Walter, H.: Geno2pheno: estimating phenotypic drug resistance from HIV-1 genotypes. NAR, 31 (2003) 3850-3855

48. Beerenwinkel, N., Schmidt, B., Walter, H., Kaiser, R., Lengauer, T., Hoffmann, D., Korn, K.,Selbig, J.,: Diversity and complexity of HIV-1 drug resistance: a bioinformatics approach to predicting phenotype from genotype. PNAS, 99 (2002) 8271-8276
49. Zazzi, M., Romano, L., Giulietta, V., Shafer, R.W., Reid, C., Bello, F., Parolin, C., Palu, G., Valensin, P.: Comparative evaluation of three computerized algorithms for prediction of antiretroviral susceptibility from HIV type 1 genotype. Journal of Antiimicrobial Chemotherapy, 53 (2004) 356-360
50. Sa-Filho, D. J., Costa, L.J., de Oliceira, C.F., Guimaraes, A.P.C., Accetturi, C.A., Tanuri, A., Diaz, R.S.: Analysis of the protease sequences of HIV-1 infected individuals after Indinavir monotherapy. Journal of Clinical Virology, 28 (2003) 186-202