Available online at www.sciencedirect.com
ScienceDirect

Pattern Recognition Letters 28 (2007) 1610–1615

Pattern Recognition Letters

www.elsevier.com/locate/patrec

Using pseudo amino acid composition to predict protein subnuclear localization: Approached with PSSM

Piyushkumar Mundra, Madhan Kumar, K. Krishna Kumar, Valadi K. Jayaraman *, Bhaskar D. Kulkarni *

Chemical Engineering and Process Development Division, National Chemical Laboratory, Dr. Homi Bhabha Road, Pune 411 008, India

Received 13 November 2006; received in revised form 21 February 2007
Available online 11 April 2007

Communicated by M. Singh

Abstract

Identification of Nuclear protein localization assumes significance as it can provide in depth insight for genome regulation and function annotation of novel proteins. A multiclass SVM classifier with various input features was employed for nuclear protein compartment identification. The input features include factor solution scores and evolutionary information (position specific scoring matrix (PSSM) score) apart from conventional dipeptide composition and pseudo amino acid composition. All the SVM classifiers with different sets of input features performed better than the previously available prediction classifiers. The jack-knife success rate thus obtained on the benchmark dataset constructed by Shen and Chou [Shen, H.B., Chou, K.C., 2005, Predicting protein subnuclear location with optimized evidence-theoretic K-nearest classifier and pseudo amino acid composition. Biochem. Biophys. Res. Commun. 337, 752–756] is 71.23%, indicating that the novel pseudo amino acid composition approach with PSSM and SVM classifier is very promising and may at least play a complimentary role to the existing methods.

Keywords: Nuclear protein; Subnuclear localization; Multiclass SVM; Factor solution score; PSSM

1. Introduction

The cell nucleus, only present in the eukaryotic cells, is a highly complex organelle. It organizes the assembly of genes and their corresponding regulatory functions. It is known that many nuclear proteins, which participate in the related nuclear function, tend to colocalize into specific areas of the nucleus (Bridger and Bickmore, 1998). Mis-localized nuclear protein can lead to human genetic disease, cancer or virally infected cells (Sutherland et al., 2001; Phair and Mistelli, 2000).

It is necessary to reveal the full complexity of the nucleus as it is directly related with the genomic regulation and function. Advancement in the experimental techniques like protein micro characterization and mass spectrometry has enabled identification of the components of large nuclear complexes. Nevertheless, with the completion of human genome project, enormous amount of the information with unknown function is available. Hence, the need to develop cheaper and faster bioinformatics tools has increased many folds. Protein subnuclear localization prediction can be potentially very useful in annotating the function of the novel proteins.

Various algorithms for protein subcellular localization prediction are available in literature (Chou and Elrod, 1999; Chou, 2000; Pan et al., 2003). Some recent methods include PLPD (Lee et al., 2006), LOCSVMPSI (Xie et al., 2005), LOC tree (Nair and Rost, 2005) PSORTb v.2.0 (Gardy et al., 2005). Protein subcellular localization prediction for human (Chou and Shen, 2006a), eukaryotes

* Corresponding authors. Tel.: +91 20 25902731 (V.K. Jayaraman); +91 20 25902610 (B.D. Kulkarni); fax: +91 20 25902612 (V.K. Jayaraman); +91 20 25902612 (B.D. Kulkarni).
E-mail addresses: vk.jayaraman@ncl.res.in (V.K. Jayaraman), bd.kulkarni@ncl.res.in (B.D. Kulkarni).

0167-8655/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2007.04.001

(Chou and Shen, 2006b), plants (Chou and Shen, 2006c), virus (Chou and Shen, 2006d) and gram negative bacteria (Chou and Shen, 2006e; Guo et al., 2006) have also been carried out. Yet, only three studies have been executed for protein subnuclear localization. These methods employ optimized evidence-theoretic KNN (Shen and Chou, 2005) and SVM classification (Lei and Dai, 2005) and gene ontology based protein similarity (Lei and Dai, 2006). Two reasons are mainly responsible for limited study in this field. These are (1) nucleus is more compact and complicated as compared to other cell compartments (Hancock, 2004). (2) Protein complexes within the cell nucleus can alter their compartments during different phases of cell cycle (Sutherland et al., 2001).

Our methodology includes the use of multiclass support vector machine (SVM) with input features extracted employing different methods. Five factor scores from a large number of amino acid indices, which reflects polarity, secondary structure, molecular volume, codon diversity and electrostatic charge, have been derived by Atchley et al. (2005). We employed these factors for extraction of the most informative features from protein sequences and use them as input feature vector for SVM based classification of the nuclear protein families. Experiments with dipeptide composition and pseudo amino acid compositions have also been performed for the sake comparison. Further, evolutionary information like PSSM score has been used for subcellular localization of the eukaryotic proteins (Xie et al., 2005). We have extended the same methodology for protein subnuclear localization prediction.

2. Materials and methods

2.1. Dataset

The accession numbers of 370 nuclear proteins was obtained from study by Shen and Chou (2005). These high quality working set were obtained by them after careful screening of a larger set of sequences (including redundancy reduction) collected from the Nuclear Protein Database (NPD) (Dellaire et al., 2003) at http://npd.hgu.mrc.ac.uk/. The sequences were retrieved from the SWISS-PROT and TREMBL data banks (Bairoch and Apweiler, 2000). Out of which, two protein sequences having protein accession number, A55311 and U43279, were not available. Finally, The working dataset consists of 368 protein sequences grouped into nine classes, viz., Cajal body [10], Chromatin [59], Heterochromatin [31], Nuclear Diffuse [64], Nuclear Pore [24], Nuclear Speckle [15], Nucleolus [115], PcG [10] and PML body [40]. The number in the square bracket represents the total number of protein in respective class.

2.2. Statistical factors

Recently, a multivariate statistical analysis on 494 amino acid attributes (obtained from online database AA-index) has been carried out to arrive at a small set of five multidimensional numerical patterns, which describe the highly interpretable covariation among the original attributes (Atchley et al., 2005). The resultant factors are linear functions of the original data that capture the underlying latent structure of the variables. These authors have further found that the transformed scores so obtained provide a general solution for a wide variety of sequence analysis problems. Table 1 corresponds to five statistical factor solution scores for each amino acid, proposed by (Atchley et al., 2005). Factor 1 reflects the simultaneous covariation in portion of exposed residues versus buried residues, polarity versus no polarity, hydrophobicity versus hydrophilicity, non-bonded energy versus free energy. This factor can be designated as polarity index. Factor 2 is a secondary structure factor, which represents the relationship of various amino acids with secondary structure configurations like helix, turn or coil. Factor 3 relates to molecular size or volume. Factor 4 reflects the relative amino acid composition in various proteins. Factor 5 refers to electrostatic charge with high coefficient on isoelectric point and net charge.

With a view to extract the most informative features, SVM classifier was built with various combinations of the five factors. For SVM based classification, it is required to convert sequences of variable length into fixed length input feature vector.

To convert the protein sequence in the fixed length vector, fraction of each amino acid in the given protein sequence is first calculated by the following equation:

$$\text{Fraction of amino acid } i = \frac{\text{Total number of amino acids of type } i}{\text{Total number of amino acids in the protein}} \qquad (1)$$

Then, factor solution scores of each amino acid were subsequently multiplied by respective amino acid fraction of

Table 1
Five factor solution score for amino acid proposed by Atchley et al. (2005)

Amino acid	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
A	−0.591	−1.302	−0.733	1.57	−0.146
C	−1.343	0.465	−0.862	−1.020	−0.255
D	1.050	0.302	−3.656	− 0.259	−3.242
E	1.357	−1.453	1.477	0.113	−0.837
F	−1.006	−0.590	1.891	− 0.397	0.412
G	−0.384	1.652	1.330	1.045	2.064
H	0.336	−0.417	−1.673	−1.474	−0.078
I	−1.239	−0.547	2.131	0.393	0.816
K	1.831	−0.561	0.533	−0.277	1.648
L	−1.019	−0.987	−1.505	1.266	−0.912
M	−0.663	−1.524	2.219	−1.005	1.212
N	0.945	0.828	1.299	−0.169	0.933
P	0.189	2.081	−1.628	0.421	−1.392
Q	0.931	−0.179	−3.005	−0.503	−1.853
R	1.538	−0.055	1.502	0.440	2.897
S	−0.228	1.399	−4.760	0.670	−2.647
T	−0.032	0.326	2.213	0.908	1.313
V	−1.337	−0.279	−0.544	1.242	−1.262
W	−0.595	0.009	0.672	−2.128	−0.184
Y	0.260	0.830	3.097	−0.838	1.512

the given protein sequence. For example, using this method and the factor solution scores shown in Table 1, for sequence 'ACEAE' the fraction of amino acids are

Fraction of 'A' = 0.4
Fraction of 'C' = 0.2
Fraction of 'E' = 0.4

The fractions of all other amino acids are zero. Multiplying these fractions with corresponding factor scores of amino acid, we obtain

Score of 'A' = [−0.2364 −0.5208 −0.2932 0.628 −0.0584]
Score of 'C' = [−0.2686 0.093 −0.1724 −0.204 −0.051]
Score of 'E' = [0.5428 −0.5812 0.5908 0.0452 −0.3348]

Scores of other amino acids are zero. Thus, the input feature vector will be [−0.2364 −0.5208 −0.2932 0.628 −0.0584 −0.2686 0.093 −0.1724 −0.204 −0.051 0 0 0 0 0 0.5428 −0.5812 0.5908 0.0452 −0.3348 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0].

As shown in the above example, all the five factor scores can be included for feature extraction. Depending upon the particular problem it may be necessary to include all the five factors or only a subset of the factors. Hence, various combinations of subsets of factors were employed for feature extraction. By this way, length of input feature vector was thus varied from 20 (with one factor solution score) to 100 (All the factor solution score) depending upon the particular subset combination employed.

2.3. Dipeptide composition

Dipeptide composition is widely used for representing the protein sequence (Chou, 1999; Liu and Chou, 1999; Kim et al., 2006; Gao et al., 2005; Idicula-Thomas et al., 2006). It converts the protein sequence into 400 input feature vectors. Dipeptide composition encapsulates the information about the fraction of amino acid as well as local order. The dipeptide composition of each protein sequence can be calculated by the following equation:

$$\text{Fraction of dipeptide } (i) = \frac{\text{Total number of dipeptide } (i) \text{ in the sequence}}{\text{Total number of all possible dipeptides}} \tag{2}$$

2.4. Pseudo amino acid composition

This method represents the protein sequence into the fraction of amino acid in protein sequence along with sequence order and length information. The first 20 features of the sequence, derived from pseudo amino acid composition method represent the composition of the twenty amino acids. Remaining components represent the sequence order effect. These features can be calculated by using physical properties of amino acids. In the present study, hydrophobicity, hydrophilicity and side chain mass of amino acids are used. These physical property values were taken from (Chou and Cai, 2006). The original values of these properties are converted into zero mean value as was done previously (Chou, 2001). Then, other elements of pseudo amino acid composition were calculated using Equations (2)–(6) of Chou (2001). Here, the sequence order effect is calculated for 14th rank (successive amino acids) ($\lambda = 14$).

2.5. Position specific scoring matrix (PSSM)

Evolutionary information of protein sequence like PSSM can be extracted from PSI-BLAST profile (Jones, 1999). This method has earlier been used for protein subcellular localization (Xie et al., 2005). Each protein sequence from all the nine classes was searched against the Non Redundant (NR) database with three iterations of PSI-BLAST with a cut off e-value of 0.001. The PSSM of a protein sequence extracted from the PSI-BLAST profile was used to generate a 400 dimensional input vector by summing up all rows in the PSSM corresponding to the original primary sequence. Every element in this input vector was subsequently divided by the length of the sequence and then scaled to the range of 0–1 by using the standard sigmoid function (Jones, 1999; Xie et al., 2005). The resultant matrix with 400 elements was used as input feature for SVM. The standard sigmoid function used was

$$\frac{1}{1+e^{-x}} \tag{3}$$

3. SVM binary classification

Support Vector Machine (SVM) is based on Vapnik's statistical learning theory (Vapnik, 1995; Muller et al., 2001). Due to its excellent generalization capabilities and ability to converge to a single globally optimal solution, it is widely used in the bioinformatics applications (Lumini and Nanni, 2006; Idicula-Thomas et al., 2006; Kulkarni et al., 2005; Rätsch et al., 2005; Lei and Dai, 2005; Zavaljevski et al., 2002; Chou and Cai, 2002).

In linearly separable problems, binary SVM determines the optimal hyperplane separates the data belonging to the two classes in such a way that the distance between the closest instances belonging to the two classes (margin) is maximized. Normally in real life problems, the data is nonlinearly separable. In such cases, in SVM classification, the input data is first mapped into a higher dimensional feature space and subsequently, a linear hyperplane is constructed to separate them. Further, to reduce the complexity, an appropriate kernel is defined such that the computations can be performed in the input space itself. Detailed discussions on SVM classification can be found in (Vapnik, 1995; Muller et al., 2001).

3.1. Multiclass SVM

Protein subnuclear localization consists of nine distinct classes. Hence, this becomes multiclass prediction problem. Normally, “One-against-one” or “One-against-all” approach is employed for multiclass SVM classifier (Hsu and Lin, 2002). In the present study, “One-against-one” approach was used. This method involves construction of individual binary SVM classifier corresponding to each pair of the classes. Hence, if there are K classes, a total of $K(K-1)/2$ classifiers will be constructed.

Unseen test instances prediction follows the voting strategy. Predictions are made with each binary classifiers and label is assigned to a class with maximum number of votes. In case when tie arise, i.e. two classes have identical votes, label assignment to the class is made on the basis of smallest index.

All the computations were performed using LIBSVM-2.81 standard package (Chang and Lin, 2001). The various user-defined parameters, e.g., kernel parameter γ and regularization parameter C were optimized on the training dataset.

3.2. SVM performance evaluation

SVM performance was evaluated using most common tests like resubstitution test and jack-knife cross-validation test.

Jack-knife test is also known as leave-one-out test. In jack-knife test, each protein in the dataset is removed and training is performed with remaining protein sequences. Thereafter, testing is done with the removed protein sequence. By this way, turn-by-turn each protein is tested for the accuracy (employing the classifier trained with the rest of the dataset) and the average accuracy value is presented. Though this method is time consuming, this is most effective, rigorous and reliable for calculating the accuracy of the classification method, especially for multi class problem.

The resubstitution test reflects the self-consistency of a classification method. Here, training dataset, which is used to generate the classification rule, is itself employed for the testing purpose. This will definitely underestimate the error and enhance the success rate because same proteins are used to construct the model and to test themselves. Though it gives the higher accuracy, this test is absolutely necessary because it represents the self-consistency of the identification method. If self-consistency is poor, that method cannot be termed as a good classification method.

4. Results and discussion

Various combinations of statistical factor scores (20–100 features), dipeptide composition (400 features), pseudo amino acid composition ($20+\lambda$ features, here $\lambda = 14$) and PSSM (400 features) were employed as input feature vector for SVM. The performance of each trained module was evaluated with resubstitution test and jack-knife cross-validation test. Resubstitution test represents the self-consistency of the prediction method. In statistical prediction, the following three cross-validation tests are often used to examine the power of a predictor: independent dataset test, sub-sampling test, and jack-knife test. Of these three, the jack-knife test is thought the most rigorous and objective one (Chou and Zhang, 1995), and hence has been used by more and more investigators (Chou and Shen, 2006a,b,c; Lei and Dai, 2005, 2006; Guo et al., 2006) in examining the power of various prediction methods.

The classification performance of different SVM modules is summarized in Table 2. Initially, all the factor solution scores were used for input feature vector representation. It resulted in 63.5% accuracy. Subsequently, to improve the jack-knife test accuracy, various combinations of the factor score were employed. Out of various combinations tried, combination of factors 1, 2 and 4 classified the proteins better than others with 66.84% jack-knife test accuracy. Dipeptide composition based module performed better with 67.39% accuracy. Further, position specific scoring matrix, which represents the evolutionary information of the protein, when used as input feature for SVM, exhibited superior performance in the jack-knife testing accuracy (71.23%). The individual accuracy for each class is summarized in Table 3.

The corresponding comparison of the other subnuclear protein localization prediction methods is shown in Table 4. All our methods, except Pseudo amino acid encoding, perform better than the results obtained by Shen and Chou.

Table 2
Prediction performance of various methods

Input features for SVM	SVM parameters		Overall success rate	
	C	γ	Resubstitution test (%)	Jack-knife test (%)
Pseudo amino acid composition with $\lambda = 14$	2000	0.6	97.28	53.26
Factors 1, 2 and 4	4	100	99.45	66.84
Dipeptide composition	5	200	99.72	67.3913
PSSM	4	20	99.17	71.23

Table 3
Individual class accuracy based on jack-knife test

Class	PSSM	Dipeptide	Factors 1,2,4	Pseudo AA
Cajal	0.2	0.2	0.2	0.1
Chromatin	0.6667	0.6949	0.661	0.6441
Heterochromatin	0.7419	0.5806	0.6452	0.5484
Nucleolus	0.9385	0.9043	0.8522	0.6
Diffuse	0.65	0.5781	0.625	0.4844
Nuclear Pore	0.5833	0.5	0.5417	0.4167
Speckle	0.5	0.4	0.333	0.2667
PcG	0.3	0.3	0.3	0.3
PML	0.6	0.575	0.6	0.525

Table 4
Comparison with other prediction methods

Algorithm	Overall success rate	
	Resubstitution test (%)	Jack-knife test (%)
OET-KNN with $\lambda = 14$[a]	97.28	64.32
SVM with amino acid composition[a]	98.92	33.78
ProtLock[a]	38.92	29.46
SVM with PSSM (present study)	99.17	71.23

[a] Results from Shen and Chou (2005).

Table 5
Individual class accuracy based on resubstitution test

Class	PSSM	Dipeptide	Factors 1,2,4	Pseudo AA
Cajal	1	1	1	1
Chromatin	1	1	1	0.97
Heterochromatin	0.97	1	1	0.97
Nucleolus	1	1	1	0.97
Diffuse	0.98	0.98	0.97	0.98
Nuclear Pore	1	1	1	1
Speckle	0.94	1	1	0.93
PcG	1	1	1	1
PML	1	1	1	0.95

Further, self-consistency of the prediction method was tested with resubstitution test. Table 1 shows the result of resubstitution test for all the methods. It clearly shows that all the prediction algorithms performed with more than 99% self-consistency except pseudo amino acid composition method. It shows that all the prediction algorithms can be termed as good classifiers. The corresponding accuracies of individual classes are summarized in Table 5.

As discussed by Atchley et al. (2005), factor 1 represent the physical properties like hydrophobicity, hydrophilicity and Polarity, factor 2 embody the secondary structure information in terms of relationship of various amino acids to helix, turn or coil, and factor 4 signifies importance of relative amino acid composition. Classification results infer that factors based on these properties can be effectively used for the prediction of subnuclear protein compartment.

Using PSSM to represent the nuclear protein is an effective way to incorporate evolutionary information. The superior performance of SVM based classifier with PSSM suggests strong correlation between evolutionary information and protein nuclear compartment. Results with the statistical factors and dipeptide composition are quite encouraging and statistical factors based feature extraction may be very relevant for similar classification tasks.

5. Conclusion

In the present study multiclass SVM, one of the most powerful classifiers, was employed for predicting protein subnuclear localization. This method gave comparable results with conventional dipeptide composition and statistical factor score based features. Further, evolutionary information, in form of PSSM score, was used as input attributes to SVM. The jack-knife success rate thus obtained on the benchmark dataset constructed by Shen and Chou is 71.23%, indicating that the novel pseudo amino acid composition approach with PSSM and SVM classifier is very promising and may at least play a complimentary role to the existing methods.

Acknowledgement

Financial assistance from Department of Biotechnology, New Delhi, is gratefully acknowledged.

References

Atchley, W.R., Zhao, J., Fernandes, A.D., Drüke, T., 2005. Solving the protein sequence metric problem. Proc. Natl. Acad. Sci. 102, 6395–6400.

Bairoch, A., Apweiler, R., 2000. The SWISS-PROT protein sequence data bank and its supplement TreEMBL. Nucleic Acid Res. 28, 45–48.

Bridger, J.M., Bickmore, W.A., 1998. Putting the genome on the map. Trends Genet. 14, 403–409.

Chang, C.C., Lin, C.J., 2001. LIBSVM: A Library for Support Vector Machines. www.csie.ntu.edu.tw/~cjlin/libsvm.

Chou, K.C., 1999. Using pair-coupled amino acid composition to predict protein secondary structure content. J. Protein Chem. 18, 473–480.

Chou, K.C., 2000. Prediction of protein subcellular locations by incorporating quasi-sequence-order effect. Biochem. Biophys. Res. Commun. 278, 477–483.

Chou, K.C., 2001. Prediction of protein cellular attributes using pseudo amino acid composition. Proteins: Struct. Funct. Genet. 43, 246–255 (Erratum: ibid., 2001, 44, 60).

Chou, K.C., Cai, Y.D., 2002. Using functional domain composition and support vector machines for prediction of protein subcellular localization. J. Biol. Chem. 277, 45765–45769.

Chou, K.C., Cai, Y.D., 2006. Prediction of protease types in a hybridization space. Biochem. Biophys. Res. Commun. 339, 1015–1020.

Chou, K.C., Elrod, D.W., 1999. Protein subcellular location prediction. Protein Eng. 12, 107–118.

Chou, K.C., Shen, H.B., 2006a. Hum-PLoc: A novel ensemble classifier for predicting human protein subcellular localization. Biochem. Biophys. Res. Commun. 347, 150–157.

Chou, K.C., Shen, H.B., 2006b. Predicting eukaryotic protein subcellular location by fusing optimized evidence-theoretic K-nearest neighbor classifiers. J. Proteome Res. 5, 1888–1897.

Chou, K.C., Shen, H.B., 2006c. Large-scale plant protein subcellular location prediction. J. Cell. Biochem. 100, 665–678.

Chou, K.C., Shen, H.B., 2006d. Virus-PLoc: A fusion classifier for predicting the subcellular localization of viral proteins within host and virus-infected cells. Biopolymers 85, 233–240.

Chou, K.C., Shen, H.B., 2006e. Large-scale predictions of Gram-negative bacterial protein subcellular locations. J. Proteome Res. 5, 3420–3428.

Chou, K.C., Zhang, C.T., 1995. Review: Prediction of protein structural classes. Crit. Rev. Biochem. Mol. Biol. 30, 275–349.

Dellaire, G., Farrall, R., Bickmore, W.A., 2003. The Nuclear Protein Database (NPD): Subnuclear localization and functional annotation of the nuclear proteome. Nucleic Acid Res. 31, 328–330.

Gao, Q.B., Wang, Z.Z., Yan, C., Du, Y.H., 2005. Prediction of protein subcellular location using a combined feature of sequence. FEBS Lett. 579, 3444–3448.

Gardy, J.L., Laird, M.R., Chen, F., Rey, S., Walsh, C.J., Ester, M., Brinkman, F.S., 2005. PSORTb v.2.0: Expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis. Bioinformatics 21, 617–623.

Guo, J., Lin, Y., Liu, X., 2006. GNBSL: A new integrative system to predict the subcellular location for Gram-negative bacteria proteins. Proteomics 6, 5099–5105.

Hancock, R., 2004. Internal organization of the nucleus: Assembly of compartments by macromolecular crowding and the nuclear matrix model. Biol. Cell 96, 595–601.

Hsu, C.W., Lin, C.J., 2002. A comparison of methods for multiclass support vector machines. IEEE Trans. Neural Networks 13, 415–425.

Idicula-Thomas, S., Kulkarni, A.J., Kulkarni, B.D., Jayaraman, V.K., Balaji, P.V., 2006. A support vector machine-based method for predicting the propensity of a protein to be soluble or to form inclusion body on overexpression in *Escherichia coli*. Bioinformatics 22, 278–284.

Jones, D.T., 1999. Protein secondary structure prediction based on position-specific scoring matrices. J. Mol. Biol. 292, 195–202.

Kim, J.K., Raghava, G.P.S., Bang, S.Y., Choi, S., 2006. Prediction of subcellular localization of proteins using pairwise sequence alignment and support vector machine. Pattern Recognition Lett. 27, 996–1001.

Kulkarni, O.C., Vigneshwar, R., Jayaraman, V.K., Kulkarni, B.D., 2005. Identification of coding and noncoding sequences using local holder exponent formalism. Bioinformatics 21, 3818–3823.

Lee, K., Kim, D.W., Na, D., Lee, K.H., Lee, D., 2006. PLPD: Reliable protein localization prediction from imbalanced and overlapped datasets. Nucleic Acids Res. 34, 4655–4666.

Lei, Z., Dai, Y., 2005a. A class of new kernels based on a matrix of high-scored pairs of k-peptides and its applications in prediction of protein sub-cellular localization. LNCS Trans. Comput. Systems Biol. II, 48–58.

Lei, Z., Dai, Y., 2005b. An SVM-based system for predicting protein subnuclear localizations. BMC Bioinformatics 6, 291–298.

Lei, Z., Dai, Y., 2006. Assessing protein similarity with Gene Ontology and its use in subnuclear localization prediction. BMC Bioinformatics 7, 491–500.

Liu, W., Chou, K.C., 1999. Protein secondary structural content prediction. Protein Eng. 12, 1041–1050.

Lumini, A., Nanni, L., 2006. Machine learning for HIV-1 protease cleavage site prediction. Pattern Recognition Lett. 27, 1537–1544.

Muller, K.R., Mika, S., Ratsch, K., Tsuda, K., Scholkopf, B., 2001. An introduction to kernel-based learning algorithms. IEEE Trans. Neural Networks 2, 181–201.

Nair, R., Rost, B., 2005. Mimicking cellular sorting improves prediction of subcellular localization. J. Mol. Biol. 348, 85–100.

Pan, Y.X., Zhang, Z.Z., Guo, Z.M., Feng, G.Y., Huang, Z.D., He, L., 2003. Application of pseudo amino acid composition for predicting protein subcellular location: Stochastic signal processing approach. J. Protein Chem. 22, 395–402.

Phair, R.D., Mistelli, T., 2000. High mobility of proteins in the mammalian cell nucleus. Nature 404, 604–609.

Rätsch, G., Sonnenburg, S., Schölkopf, B., 2005. RASE: Recognition of alternatively Spliced Exons in *C. elegans*. Bioinformatics 21, i369–i377.

Shen, H.B., Chou, K.C., 2005. Predicting protein subnuclear location with optimized evidence-theoretic K-nearest classifier and pseudo amino acid composition. Biochem. Biophys. Res. Commun. 337, 752–756.

Sutherland, H.E., Mumford, G.K., Newton, K., Ford, L.V., Farrall, R., Dellaire, G., Cáceres, J.F., Bickmore, W.A., 2001. Large-scale identification of mammalian proteins localized to nuclear sub-compartments. Hum. Mol. Genet. 10, 1995–2011.

Vapnik, V., 1995. The Nature of Statistical Learning Theory. Springer, NY.

Xie, D., Li, A., Wang, M., Fan, Z., Feng, H., 2005. LOCSVMPSI: A web server for subcellular localization of eukaryotic proteins using SVM and profile of PSI-BLAST. Nucleic Acids Res. 33, W105–W110, Web Server.

Zavaljevski, N., Stevens, F.J., Reifman, J., 2002. Support vector machines with selective kernel scaling for protein classification and identification of key amino acid positions. Bioinformatics 18, 689–696.