Protein Engineering, Design & Selection vol. 20 no. 11 pp. 561–567, 2007
Published online November 10, 2007 doi:10.1093/protein/gzm057

Nuc-PLoc: a new web-server for predicting protein subnuclear localization by fusing PseAA composition and PsePSSM

Hong-Bin Shen[1,2,4,5] and Kuo-Chen Chou[1,3]

[1]Institute of Image Processing and Pattern Recognition, Shanghai Jiaotong University, 1954 Hua-Shan Road, Shanghai 200030, China, [2]School of Information Engineering, Jiangnan University, Wuxi 214122 and [3]Gordon Life Science Institute, San Diego, CA 92130, USA

[4]Present address: Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, MA 02115, USA

[5]To whom correspondence should be addressed.
E-mail: hbshen@crystal.harvard.edu

The life processes of an eukaryotic cell are guided by its nucleus. In addition to the genetic material, the cellular nucleus contains many proteins located at its different compartments, called subnuclear locations. Information of their localization in a nucleus is indispensable for the in-depth study of system biology because, in addition to helping determine their functions, it can provide illuminative insights of how and in what kind of microenvironments these subnuclear proteins are interacting with each other and with other molecules. Facing the deluge of protein sequences generated in the post-genomic age, we are challenged to develop an automated method for fast and effectively annotating the subnuclear locations of numerous newly found nuclear protein sequences. In view of this, a new classifier, called Nuc-PLoc, has been developed that can be used to identify nuclear proteins among the following nine subnuclear locations: (1) chromatin, (2) heterochromatin, (3) nuclear envelope, (4) nuclear matrix, (5) nuclear pore complex, (6) nuclear speckle, (7) nucleolus, (8) nucleoplasm and (9) nuclear promyelocytic leukaemia (PML) body. Nuc-PLoc is featured by an ensemble classifier formed by fusing the evolution information of a protein and its pseudo-amino acid composition. The overall jackknife cross-validation accuracy obtained by Nuc-PLoc is significantly higher than those by the existing methods on the same benchmark data set through the same testing procedure. As a user-friendly web-server, Nuc-PLoc is freely accessible to the public at http://chou.med.harvard.edu/bioinf/Nuc-PLoc.

Keywords: fusion/Nuc-PLoc/position-specific scoring matrix/pseudo-amino acid composition/subnuclear location

Introduction

The nucleus exists only in eukaryotic cells. Located at the center of a cell like its kernel, the nucleus is the most prominent and largest cellular organelle (Lodish *et al*., 1995), with the diameter from 11 to 22 μm and occupying ~10% of the total volume of a typical animal cell (Alberts *et al*., 2002). Similar to a cell filled with the cytoplasm, a nucleus is also filled with a viscous liquid called nucleoplasm. Again, similar to a cell enclosed by the cell membrane, a nucleus is also surrounded by a membrane, called the nuclear envelope. The envelope is riddled with holes, called nuclear pores, to allow specific materials to pass in and out of the nucleus, just like proteins in the cell membrane that regulate the movement of molecules in and out of the cell itself.

Functioning as the 'brain' of eukaryotic cells, the nucleus guides the life processes of the cells by directing their reproduction, controlling their differentiation and regulating their metabolic activities. In addition to the genetic material, a nucleus contains many proteins located at its different compartments, called subnuclear locations. Information of the subnuclear locations of these proteins is important because it not only provides useful clues about their functions but also helps understand how and in what kind of microenvironments they interact with each other and with other molecules, and hence is indispensable for the in-depth study of system biology at the cell nucleus level.

Although the protein's subnuclear localization can be determined by conducting various experiments, such as cell fractionation, electron microscopy and fluorescence microscopy (Murphy *et al*., 2000), it is both time-consuming and costly to acquire such information solely by experiments. With the deluge of protein sequences generated in the post-genomic age, it is highly desired to develop an automated method for efficiently identifying the subnuclear location of a query protein according to its sequence. Actually, many methods have been proposed for predicting protein subcellular localization (see, e.g., Nakai and Kanehisa, 1992; Nakashima and Nishikawa, 1994; Cedano *et al*., 1997; Chou and Elrod, 1999; Nakai and Horton, 1999; Yuan, 1999; Emanuelsson *et al*., 2000, 2007; Nakai, 2000; Feng, 2001, 2002; Feng and Zhang, 2001; Hua and Sun, 2001; Chou and Cai, 2002; Nair and Rost, 2002; Gardy *et al*., 2003; Pan *et al*., 2003; Park and Kanehisa, 2003; Zhou and Doctor, 2003; Huang and Li, 2004; Gao *et al*., 2005; Garg *et al*., 2005; Lei and Dai, 2005; Matsuda *et al*., 2005; Xiao *et al*., 2005; Guo *et al*., 2006; Hoglund *et al*., 2006; Lee *et al*., 2006; Pierleoni *et al*., 2006; Zhang *et al*., 2006b, 2006c; Chou and Shen, 2007a; Shen and Chou, 2007; Shi *et al*., 2007; and the references cited in a recent review (Chou and Shen, 2007b)); in contrast, however, much fewer prediction methods (particularly with web-server) have been reported for predicting the protein subnuclear localization (Lei and Dai, 2005; Shen and Chou, 2005a). The present study was initiated in an attempt to enrich the latter by introducing a novel and powerful approach through fusing the pseudo-amino acid composition (Chou, 2001) and position-specific scoring matrix (Altschul *et al*., 1997), in hope to stimulate the development of this area, which is vitally important for in-depth understanding of the biological pathways in nucleus.

Materials and methods

Protein sequences were collected from the Swiss-Prot database (version 52.0 released on 6 May 2007) at

http://www.ebi.ac.uk/swissprot/ according to the annotation information in the CC (comment or notes) field. In order to collect as much desired information as possible, but meanwhile ensure a high quality for the working data sets, the data were screened strictly according to the following criteria. (1) Because a same subnuclear location (-!-SUBCELLULAR LOCATION) in the CC field might be annotated with different terms, several key words were used for a same subcellular location. For example, in search for nuclear envelope proteins, the key words 'nuclear envelope', 'nuclear inner membrane' and 'nuclear outer membrane' were used. (2) Sequences annotated with ambiguous or uncertain terms, such as 'potential', 'probable', 'probably', 'maybe', 'likely' or 'by similarity', were excluded. (3) Sequences annotated by two or more locations were not included because of lack of the uniqueness. (4) Sequences annotated with 'fragment' were excluded; also, sequences with <50 amino acid residues were removed because they might just be fragments. (5) To avoid any homology bias, a redundancy cutoff was operated by a culling program to winnow those sequences which have $\leq 80\%$ sequence identity to any other in a same subnuclear location.

After strictly following the above five procedures, we obtained 714 proteins, of which 99 belong to chromatin, 22 to heterochromatin, 61 to nuclear envelope, 29 to nuclear matrix, 79 to nuclear pore complex, 67 to nuclear speckle, 307 to nucleolus, 37 to nucleoplasm and 13 to nuclear PML body (Fig. 1). Each of the nine subnuclear locations corresponds to a subset $\mathbb{S}_i$ ($i = 1, 2, \ldots, 9$) as shown in Table I. Thus, the benchmark data set $\mathbb{S}$ is a union of nine subsets, i.e.

$$\mathbb{S} = \mathbb{S}_1 \cup \mathbb{S}_2 \cup \mathbb{S}_3 \cup \mathbb{S}_4 \cup \mathbb{S}_5 \cup \mathbb{S}_6 \cup \mathbb{S}_7 \cup \mathbb{S}_8 \cup \mathbb{S}_9 \quad (1)$$

where $\cup$ is the symbol for union in the set theory. The sequences of the 714 subnuclear proteins as well as their accession numbers are given in Online Supporting Information A available at *PEDS* online.

It is instructive to point out that the benchmark data set constructed here is different from that of Shen and Chou (2005a). The reasons for us to re-construct the benchmark data set are as follows: (i) much more nuclear protein data are available now in Swiss-Prot database that allows us to construct a benchmark data set with a higher quality and (ii) the sequences in the original data set (Shen and Chou, 2005a) were not treated by a cutoff procedure as done here to reduce the redundancy and homologous bias.

To represent a protein sample $\mathbf{P}$ with L amino acid residues by its evolution information, the position-specific scoring matrix (PSSM) was introduced as its descriptor, i.e.

$$\mathbf{P}_{\mathrm{PSSM}} = \begin{bmatrix} \mathbb{M}_{1\to1} & \mathbb{M}_{1\to2} & \ldots & \mathbb{M}_{1\to20} \\ \mathbb{M}_{2\to1} & \mathbb{M}_{2\to2} & \ldots & \mathbb{M}_{2\to20} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbb{M}_{i\to1} & \mathbb{M}_{i\to2} & \ldots & \mathbb{M}_{i\to20} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbb{M}_{L\to1} & \mathbb{M}_{L\to2} & \ldots & \mathbb{M}_{L\to20} \end{bmatrix} \quad (2)$$

where $\mathbb{M}_{i\to j}$ represents the score of the amino acid residue in the ith position of the protein sequence being mutated to amino acid type j during the evolution process. Here, for simplifying the formulation without losing generality, let us use the numerical codes $1, 2, \ldots, 20$ to represent the 20 native amino acid types according to the alphabetical order of their single character codes. The $L \times 20$ scores in the matrix of Eq.(2) for $\mathbf{P}_{\mathrm{PSSM}}$ were generated using PSI-BLAST (Schaffer *et al.*, 2001) to search the Swiss-Prot database (version 52.0, released on 6 May 2007) for multiple sequence alignment against the protein sample $\mathbf{P}$, followed

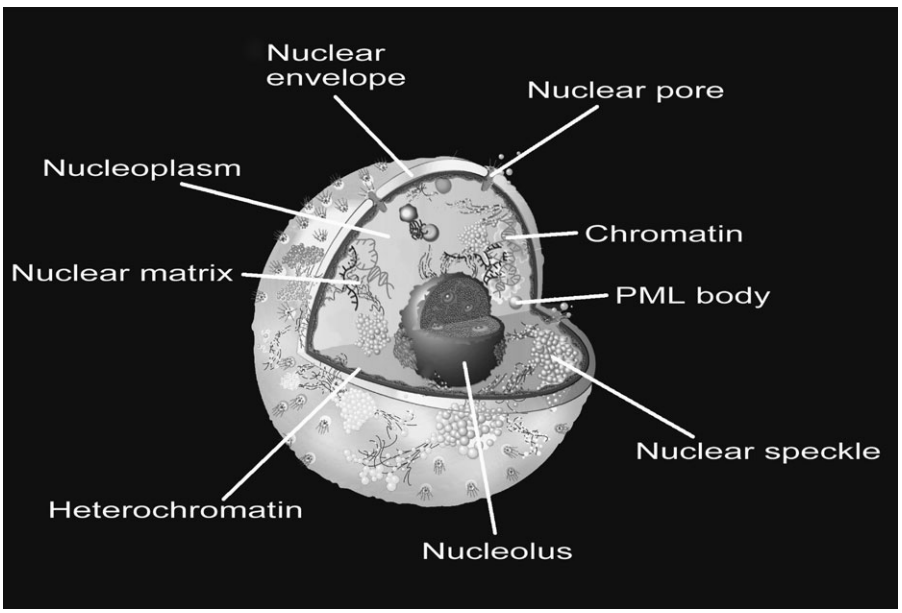

Fig. 1. Schematic drawing to show the nine subnuclear locations: (1) chromatin, (2) heterochromatin, (3) nuclear envelope, (4) nuclear matrix, (5) nuclear pore complex, (6) nuclear speckle, (7) nucleolus, (8) nucleoplasm, and (9) nuclear PML body. Adapted from Spector (2001) with permission.

Table I. The benchmark data set consists of 714 nuclear proteins classified into nine subnuclear locations (Fig. 1)

Subnuclear location	Subset[a]	Number of proteins[b]
Chromatin	$\mathbb{S}_1$	99
Heterochromatin	$\mathbb{S}_2$	22
Nuclear envelope	$\mathbb{S}_3$	61
Nuclear matrix	$\mathbb{S}_4$	29
Nuclear pore complex	$\mathbb{S}_5$	79
Nuclear speckle	$\mathbb{S}_6$	67
Nucleolus	$\mathbb{S}_7$	307
Nucleoplasm	$\mathbb{S}_8$	37
Nuclear PML body	$\mathbb{S}_9$	13
Overall	$\mathbb{S}$	714

[a]See Eq. (1).
[b]The protein sequences are given in Online Supporting Information A available at *PEDS* online.

by a standardization procedure given below.

$$\mathbb{M}_{i\to j}=\frac{\mathbb{M}^0_{i\to j}-1/20\sum_{k=1}^{20}\mathbb{M}^0_{i\to k}}{\sqrt{\frac{1}{20}\sum_{u=1}^{20}\left(\mathbb{M}^0_{i\to u}-1/20\sum_{k=1}^{20}\mathbb{M}^0_{i\to k}\right)^2}} \quad (i=1,2,\ldots,L; j=1,2,\ldots,20) \tag{3}$$

where $\mathbb{M}^0_{i\to j}$ represents the original scores directly created by PSI-BLAST that are generally shown as positive or negative integers. This is not the case for the converted scores, which will have a zero mean value over the 20 amino acids and will remained unchanged if going through the same conversion procedure again. The positive score means that the corresponding mutation occurs more frequently in the alignment than expected by chance, whereas the negative score means just the opposite. Large positive scores often indicate critical functional residues, such as active site residues and residues required for interactions with other molecules. However, according to the PSSM descriptor [Eq. (2)], proteins with different lengths will correspond to matrices with different numbers of rows. To make the PSSM descriptor become a uniform representation, one possible approach is to represent a protein sample **P** by

$$\overline{\mathbf{P}}_{\mathrm{PSSM}}=[\overline{\mathbb{M}}_1\ \ \overline{\mathbb{M}}_2\ \ \ldots\ \overline{\mathbb{M}}_{20}]^{\mathrm{T}} \tag{4}$$

where **T** is the transpose operator, and

$$\overline{\mathbb{M}}_j=\frac{1}{L}\sum_{i=1}^{L}\mathbb{M}_{i\to j}\quad (j=1,2,\ldots,20) \tag{5}$$

where $\overline{\mathbb{M}}j$ represents the average score of the amino acid residues in the protein **P** being mutated to amino acid type j during the evolution process. However, if $\overline{\mathbf{P}}_{\mathrm{PSSM}}$ of Eq. (4) was used to represent the protein **P**, all the sequence-order information would be lost. To avoid complete loss of the sequence-order information, the concept of the pseudo-amino acid composition as originally proposed in Chou (2001) was adopted, i.e. instead of Eq. (4), let us use the pseudo-position-specific scoring matrix (PsePSSM) as given by

$$\overline{\mathbf{P}}^{\xi}_{\mathrm{PsePSSM}}=[\overline{\mathbb{M}}_1\ \ \overline{\mathbb{M}}_2\ \ \ldots\ \overline{\mathbb{M}}_{20}\ \ G^{\xi}_1\ \ G^{\xi}_2\ \ \ldots\ \ G^{\xi}_{20}]^{\mathbf{T}} \tag{6}$$

to represent the protein **P**, where

$$G^{\xi}_j=\frac{1}{L-\xi}\sum_{i=1}^{L-\xi}[\mathbb{M}_{i\to j}-\mathbb{M}_{(i+\xi)\to j}]^2\quad (j=1,2,\ldots,20;\xi<L) \tag{7}$$

meaning that G^1_j is the correlation factor by coupling the most contiguous PSSM scores along the protein chain for the amino acid type j; G^2_j that by coupling the second-most contiguous PSSM scores; and so forth. Although the value allowed for ξ can be 0, 1, 2, ..., or 49, preliminary test results indicated that when $\xi>10$, the corresponding success rate dropped down. To simplify the problem, we can just focus on the optimal region of $\xi=0$, 1, ..., and 10. When $\xi=0$, Eq. (6) is degenerated to Eq. (4).

On the other hand, according to the representation of the pseudo-amino acid composition (PseAA) as defined in Chou (2001), the protein **P** is formulated by

$$\mathbf{P}^{\lambda}_{\mathrm{PseAA}}=[p_1\ \ p_2\ \ \ldots\ \ p_{20}\ \ p_{20+1}\ \ p_{20+2}\ \ \ldots\ \ p_{20+\lambda}]^{\mathbf{T}}\quad (\lambda<L) \tag{8}$$

where p_1, p_2, ..., p_{20} are associated with the conventional amino acid composition reflecting the occurrence frequencies of the 20 native amino acids in the protein **P** (Nakashima *et al*., 1986; Chou and Zhang, 1994), whereas the remaining components p_{20+1}, p_{20+2}, ..., $p_{20+\lambda}$ are the λ correlation factors that reflect the first tier, second tier, ..., and the λth tier sequence order correlation patterns, respectively (see Fig. 1 of Chou, 2001). For a given protein sequence, the $20+\lambda$ elements in Eq. (8) can be easily derived by the PseAAC web-server at http://chou.med.harvard.edu/bioinf/PseAAC/ or by Eqs (2–6) of Chou (2001). It is the additional λ factors that approximately incorporate the sequence-order effects. In this study, the optimal range for λ is from 1 to 20. Using the PseAA composition descriptor to represent protein samples as such can significantly improve the prediction quality for the subcellular localization of proteins and their other attributes as demonstrated by a series of recent publications (Pan *et al*., 2003; Chen *et al*., 2006a, 2006b; Du and Li, 2006; Mondal *et al*., 2006; Zhang *et al*., 2006a; Chen and Li, 2007; Kurgan *et al*., 2007; Lin and Li, 2007a, 2007b; Pu *et al*., 2007; Shi *et al*., 2007).

According to the PsePSSM descriptor [Eq. (6)], a protein can be represented by 11 different vectors, each of which corresponds to a different ξ (0, 1, ..., or 10), whereas according to the PseAA composition descriptor [Eq. (8)], it can be represented by 20 different vectors, each of which corresponds to a different λ (1, 2, ..., or 20). To avoid the over-fitting problem and reducing the cluster-tolerance capacity, instead of using a higher dimensional vector to represent the protein by combining the $11+20=31$ vectors of different ξ and λ, we are to introduce 31 individual basic classifiers each of which is trained and operated based on one of the aforementioned 31 descriptors. The final result is determined by an ensemble classifier formed by fusing the

31 basic classifiers through a voting system, as will be detailed below.

For the convenience of the later formulation, let us use the following equation to cover both $\overline{\mathbf{P}}^{\xi}_{\mathrm{PsePSSM}}$ and $\mathbf{P}^{\lambda}_{\mathrm{PseAA}}$ for representing a protein sample:

$$\mathbf{P}(\Phi) = \begin{cases} \overline{\mathbf{P}}^{\Phi-1}_{\mathrm{PsePSSM}}, & \text{if } \Phi = 1, 2, \ldots, 11 \\ \overline{\mathbf{P}}^{\Phi-11}_{\mathrm{PseAA}}, & \text{if } \Phi = 12, 13, \ldots, 31 \end{cases} \quad (9)$$

In this study, the optimized evidence-theoretic K nearest neighbor (OET-KNN) classifier was utilized to identify the subnuclear location of a query protein. The OET-KNN classifier is a very powerful classification engine as demonstrated by its role in enhancing the success rates of predicting membrane types (Shen and Chou, 2005b), where a detailed formulation of OET-KNN classifier can be found. There are two parameters that may directly affect the predicted result of an OET-KNN classifier. One is K, the number of the nearest proteins counted against the query protein during the prediction process; the other is Φ, i.e. which of the 31 descriptors in Eq. (9) is used as the base of the classifier. Accordingly, here the OET-KNN classifier should be formulated as an operator with the parameters K and Φ explicitly shown, i.e.

$$\text{OET - KNN} = \mathbb{C}(K, \Phi) \quad (10)$$

implying that the predicted result will depend on how to choose K and Φ. It is time-consuming and tedious to test the results using different numbers of K and Φ one by one in order for getting the optimal result. To solve such a problem, the following two-dimensional fusion approach was adopted. Preliminary tests indicated that the success rates obtained by $\mathbb{C}(K, \Phi)$ trained by the current benchmark data set became remarkably lower when $K > 10$, so it is sufficient to just consider:

$$K \in \{1, 2, \ldots, 10\}; \quad \Phi \in \{1, 2, \ldots, 31\} \quad (11)$$

where $\in$ is a symbol in the set theory meaning 'member of', then we have a set of $10 \times 31 = 310$ individual classifiers as expressed by

$$\mathbb{C}(K, \Phi), (K = 1, 2, \ldots, 10; \Phi = 1, 2, \ldots, 31) \quad (12)$$

where $\mathbb{C}(1, 1)$ is the OET-KNN classifier trained according to the 1-nearest-neighbor rule in the degenerated 20-D PSSM space [cf. Eq. (6)], $\mathbb{C}(2, 2)$ is the classifier trained according to the 2-nearest-neighbor rule in the 40-D PsePSSM space with $\xi = 2$, and so forth. The ensemble classifier formed by fusing such 310 individual classifiers is formulated by

$$\langle \mathbb{C} \rangle = \forall_{K=1}^{10} \forall_{\Phi=1}^{31} \mathbb{C}(K, \Phi) \quad (13)$$

where the symbol $\forall$ denotes the fusion operator. The detailed process of how the ensemble classifier $\langle \mathbb{C} \rangle$ works is as follows. Suppose the predicted classification result by $\mathbb{C}(K, \Phi)$ for the query protein $\mathbf{P}$ is

$$\mathbb{C}(K, \Phi) \triangleright \mathbf{P} = C_{K,\Phi} \in \mathbb{S} \quad (14)$$

where $\triangleright$ is the action operator with the meaning of using $\mathbb{C}(K, \Phi)$ to identify $\mathbf{P}$, leading to the result of $C_{K,\Phi}$ which is a member of $\mathbb{S}$ as defined by Eq. (1). The voting score for the query protein $\mathbf{P}$ belonging to the ith subset (subnuclear location) $\mathbb{S}_i$ is given by

$$\mathrm{Q}_i = \sum_{K=1}^{10} \sum_{\Phi=1}^{31} w_{K,\Phi} \Delta(C_{K,\Phi}, \mathbb{S}_i), \quad (i = 1, 2, \ldots, 9) \quad (15)$$

where $w_{K,\Phi}$ is the weight and was set at 1 for simplicity, the delta function in Eq. (15) is given by

$$\Delta(C_{K,\Phi}, \mathbb{S}_i) = \begin{cases} 1, \text{if } C_{K,\Phi} \in \mathbb{S}_i \\ 0, \text{otherwise} \end{cases}, \quad (i = 1, 2, \ldots, 9) \quad (16)$$

thus, the query protein $\mathbf{P}$ is predicted belonging to the subset for which the score of Eq. (15) is the highest, i.e.

$$\mu = \mathbf{arg\,max}_i \{\mathrm{Q}_i\}, \quad (i = 1, 2, \ldots, 9) \quad (17)$$

where μ is the argument of i that maximizes Q_i. If there is a tie among two or more subsets, then the final predicted subnuclear location will be randomly assigned to one of their corresponding subsets, although this kind of tie case rarely happens and actually was not observed in the current study. In reality, such a tie case could be very interesting because it indicates that the query protein might simultaneously exist at two or more subnuclear locations and such proteins are particularly intriguing to both basic research and drug discovery (Chou and Shen, 2007; Shen and Chou, 2007).

To provide an intuitive picture, a flowchart is provided in Fig. 2 to show the process of how the ensemble classifier works in identifying protein subnuclear localization.

Results and discussion

As a demonstration, the jackknife test was performed with the current approach on the benchmark data set (see Table I and Online Supporting Information A available at *PEDS* online). The jackknife test is deemed the most objective and rigorous cross-validation procedure in statistical prediction (Chou and Zhang, 1995) as analyzed in a recent review (Chou and Shen, 2007b) and has been used increasingly by investigators (Zhou, 1998; Pan *et al*., 2003; Zhou and Doctor, 2003; Huang and Li, 2004; Cao *et al*., 2006; Chen *et al*., 2006a, 2006b; Du and Li, 2006; Gao and Wang, 2006; Gao *et al*., 2005; Guo *et al*., 2006; Kedarisetti *et al*., 2006; Mondal *et al*., 2006; Zhang *et al*., 2006a, 2006c; Jahandideh *et al*., 2007; Lin and Li, 2007a, 2007b; Shi *et al*., 2007) to examine the power of various prediction methods.

The results thus obtained are given in Table II, where for facilitating comparison, the corresponding results by ProtLoc (Cedano *et al*., 1997), support vector machines (SVM) (Vapnik, 1998) and single OET-KNN classifier (Shen and Chou, 2005a) are also listed.

The 20-D amino acid composition (a special case of PseAA composition when $\lambda = 0$) was widely used to represent the protein samples in bioinformatics for predicting various attributes of proteins. However, the 20-D amino acid composition does not contain any sequence order information. To avoid completely losing the sequence order information, the PseAA composition [Eq. (8)] was proposed by Chou (2001). As can be seen from Table II, the prediction accuracy based on PseAA composition ($\lambda = 14$) is 7–19% higher than those based on the conventional 20-D amino acid composition.

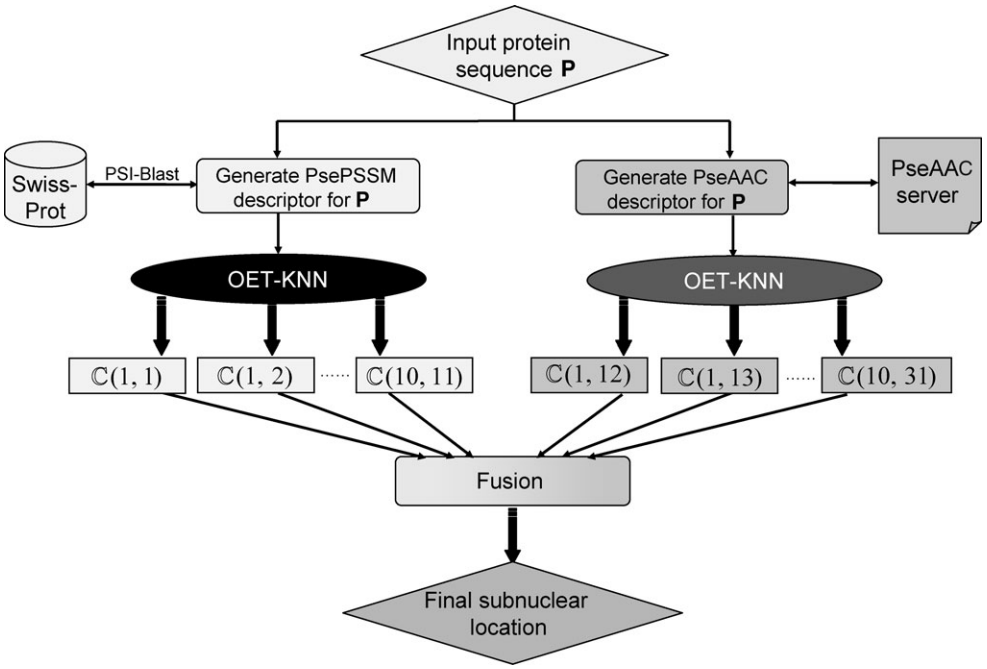

Fig. 2. A flowchart to show the process of how the ensemble classifier works by combining the protein evolution information and pseudo-amino acid composition information to identify protein subnuclear localization.

Also, as shown in Table II, the overall success rate by the jackknife test obtained with the current ensemble classifier by fusing PsePSSM and PseAA is 67.4%, which is ~31% and 19% higher than the rates obtained by ProtLoc (Cedano *et al*., 1997) and SVM (Vapnik, 1998) based on the conventional amino acid composition, and ~12% higher than the rate obtained by the single OET-KNN classifier based on PseAA composition (Shen and Chou, 2005a). The SVM predictor used in this study was C-SVC type and was trained based on radial basis function (RBF) kernel function with the parameter of $\gamma = 0.5$.

In order to demonstrate the power of the ensemble classifier formulated in this paper, we also compare the performance of a single base classifier with the ensemble classifier. It was observed that if the prediction was conducted in the PseAA composition space, the success rate obtained by the ensemble classifier was 10–13% higher than those by the individual classifiers, and that, if the prediction was conducted in the PsePSSM space, the ensemble classifier was superior to the individual classifiers by 3–7%. All these evidences indicate that the predictions obtained by individual classifiers might cause bias, prone to lead to false results.

Listed in Table III are the Matthew's correlation coefficient (MCC) indexes for the nine subnuclear locations obtained by the jackknife tests with the SVM algorithm and the current predictor, respectively. The definition of MCC is given by

$$\mathrm{MCC} = \frac{(\mathrm{TP})(\mathrm{TN}) - (\mathrm{FP})(\mathrm{FN})}{\sqrt{[\mathrm{TP}+\mathrm{FP}][\mathrm{TP}+\mathrm{FN}][\mathrm{TN}+\mathrm{FP}][\mathrm{TN}+\mathrm{FN}]}} \qquad (18)$$

Table II. Overall success rates by jackknife tests with different algorithms on the benchmark data set of Table I and Online Supporting Information A available at *PEDS* online

Algorithm	Protein sample descriptor	Overall jackknife success rate
ProtLoc[a]	Amino acid composition[b]	261/714 = 36.6%
SVM[c]	Amino acid composition[b]	349/714 = 48.9%
OET-KNN[d]	PseAA Composition[e]	397/714 = 55.6%
Ensemble classifier[f]	Fusion of PsePSSM and PseAA Composition[g]	481/714 = 67.4%

[a]See Cedano *et al*. (1997).
[b]Corresponding to $\mathbf{P}^{\lambda}_{\mathrm{PseAA}}$ of Eq. (8) with $\lambda = 0$.
[c]See Vapnik (1998).
[d]See Shen and Chou (2005a).
[e]Corresponding to $\mathbf{P}^{\lambda}_{\mathrm{PseAA}}$ of Eq. (8) with $\lambda = 14$.
[f]See $\langle \mathbb{C} \rangle$ of Eq. (13).
[g]Corresponding to $\mathbf{P}(\Phi)$ of Eq. (9) with $\Phi = 1, 2, \ldots,$ and 31.

Table III. The MCC indexes obtained by the jackknife tests with the current ensemble classifier and the SVM approach on the benchmark data set of Table I and Online Supporting Information A available at *PEDS* online

Subnuclear location	Matthew's correlation coefficient	
	SVM[a]	Ensemble classifier[a]
Chromatin $\mathbb{S}_1$	0.08	0.60
Heterochromatin $\mathbb{S}_2$	0/0[b]	0.52
Nuclear envelope $\mathbb{S}_3$	0.16	0.53
Nuclear matrix $\mathbb{S}_4$	0/0[b]	0.52
Nuclear pore complex $\mathbb{S}_5$	0.41	0.70
Nuclear speckle $\mathbb{S}_6$	0.26	0.43
Nucleolus $\mathbb{S}_7$	0.29	0.57
Nucleoplasm $\mathbb{S}_8$	0/0[b]	0.31
Nuclear PML body $\mathbb{S}_9$	0/0[b]	0.32

[a]See Table II for further explanation.
[b]The entry has no definition because both the numerator and the denominator are zero. This kind of singular case occurred when using SVM to perform the jackknife test for small subsets, such as $\mathbb{S}_2$, $\mathbb{S}_4$, $\mathbb{S}_8$, and $\mathbb{S}_9$ (Table I).

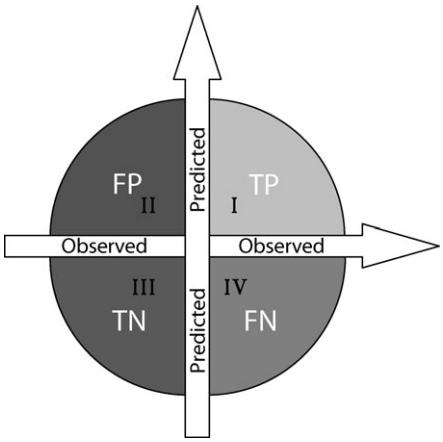

Fig. 3. An illustration to show (I) TP quadrant (green) for correct prediction of positive data set, (II) FP quadrant (red) for incorrect prediction of negative data set; (III) TN quadrant (blue) for correct prediction of negative data set; and (IV) FN quadrant (pink) for incorrect prediction of positive data set. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.).

where TP represents the true positive; TN, the true negative; FP, the false positive and FN, the false negative (Fig. 3). It can be seen from Table III that the results obtained by the current predictor not only possess higher success rates but also are more stable than those by the SVM approach, indicating that the new approach is indeed very powerful and promising.

Nuc-PLoc server is implemented with C language and HTML programming in Fedora Linux system and can be accessed freely at http://chou.med.harvard.edu/bioinf/Nuc-PLoc. On the basis of the locally computation under the configuration of AMD Athlon(tm) dual core processor 4200+ and 2.0G RAM memory, one can obtain the prediction result in 30 ± 12 seconds for each query sequence.

Conclusion

The following conclusions have been drawn through this study. (i) The success rate in identifying the protein subnuclear localization can be significantly enhanced by incorporating the protein evolution information. (ii) The ensemble classifier formed by fusing a series of basic classifiers through a voting system is a very efficient approach that allows the predictor to cover as much information as possible without causing the over-fitting problem.

To support the people working in the relevant area, a webserver called Nuc-PLoc is provided at http://chou.med.harvard.edu/bioinf/Nuc-PLoc, which is freely accessible to the public.

Acknowledgements

The authors wish to express their gratitude to the two anonymous reviewers, whose constructive comments were very helpful in strengthening the presentation of this study.

References

Alberts,B., Johnson,A., Lewis,J., Raff,M., Roberts,K. and Walter,P. (2002) *Molecular Biology of the Cell*, 4th edn. Garland Science, New York.
Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) *Nucleic Acids Res.*, **25**, 3389–3402.
Cao,Y., Liu,S., Zhang,L., Qin,J., Wang,J. and Tang,K. (2006) *BMC Bioinformatics*, **7**, 20.
Cedano,J., Aloy,P., P'erez-Pons,J.A. and Querol,E. (1997) *J. Mol. Biol.*, **266**, 594–600.
Chen,C., Tian,Y.X., Zou,X.Y., Cai,P.X. and Mo,J.Y. (2006a) *J Theor Biol.*, **243**, 444–448.
Chen,C., Zhou,X., Tian,Y., Zou,X. and Cai,P. (2006b) *Anal Biochem.*, **357**, 116–121.
Chen,Y.L. and Li,Q.Z. (2007) *J. Theor. Biol.*, doi:10.1016/j.jtbi.2007.05.019.
Chou,K.C. (2001) *Proteins Struct. Funct. Genet.*, **43**, 246–255. (Erratum: ibid., 2001, **44**, 60).
Chou,K.C. and Cai,Y.D. (2002) *J. Biol. Chem.*, **277**, 45765–45769.
Chou,K.C. and Elrod,D.W. (1999) *Protein Eng.*, **12**, 107–118.
Chou,K.C. and Shen,H.B. (2007a) *J. Proteome Res.*, **6**, 1728–1734.
Chou,K.C. and Shen,H.B. (2007b) *Anal Biochem.*, **370**, 1–16.
Chou,K.C. and Zhang,C.T. (1994) *J. Biol. Chem.*, **269**, 22014–22020.
Chou,K.C. and Zhang,C.T. (1995) *Crit. Rev. Biochem. Mol. Biol.*, **30**, 275–349.
Du,P. and Li,Y. (2006) *BMC Bioinformatics*, **7**, 518.
Emanuelsson,O., Brunak,S., von Heijne,G. and Nielsen,H. (2007) *Nat Protoc.*, **2**, 953–971.
Emanuelsson,O., Nielsen,H., Brunak,S. and von Heijne,G. (2000) *J. Mol. Biol.*, **300**, 1005–1016.
Feng,Z.P. (2001) *Biopolymers*, **58**, 491–499.
Feng,Z.P. (2002) *In Silico Biol.*, **2**, 291–303.
Feng,Z.P. and Zhang,C.T. (2001) *Int. J. Biol. Macromol.*, **28**, 255–261.
Gao,Q.B. and Wang,Z.Z. (2006) *Protein Eng. Des. Sel.*, **19**, 511–516.
Gao,Q.B., Wang,Z.Z., Yan,C. and Du,Y.H. (2005) *FEBS Lett.*, **579**, 3444–3448.
Gardy,J.L., *et al*. (2003) *Nucleic Acids Res.*, **31**, 3613–3617.
Garg,A., Bhasin,M. and Raghava,G.P. (2005) *J. Biol. Chem.*, **280**, 14427–14432.
Guo,J., Lin,Y. and Liu,X. (2006) *Proteomics*, **6**, 5099–5105.
Hoglund,A., Donnes,P., Blum,T., Adolph,H.W. and Kohlbacher,O. (2006) *Bioinformatics*, **22**, 1158–1165.
Hua,S. and Sun,Z. (2001) *Bioinformatics*, **17**, 721–728.
Huang,Y. and Li,Y. (2004) *Bioinformatics*, **20**, 21–28.
Jahandideh,S., Abdolmaleki,P., Jahandideh,M. and Asadabadi,E.B. (2007) *Biophys. Chem.*, **128**, 87–93.
Kedarisetti,K.D., Kurgan,L.A. and Dick,S. (2006) *Biochem. Biophys. Res. Commun.*, **348**, 981–988.
Kurgan,L.A., Stach,W. and Ruan,J. (2007) *J. Theor. Biol.*, doi.org/10.1016/j.jtbi.2007.05.017.
Lee,K., Kim,D.W., Na,D., Lee,K.H. and Lee,D. (2006) *Nucleic Acids Res.*, **34**, 4655–4666.
Lei,Z. and Dai,Y. (2005) *BMC Bioinformatics*, **6**, 291.
Lin,H. and Li,Q.Z. (2007a) *Biochem. Biophys. Res. Commun.*, **354**, 548–551.
Lin,H. and Li,Q.Z. (2007b) *J. Comput. Chem.*, **28**, 1463–1466.
Lodish,H., Baltimore,D., Berk,A., Zipursky,S.L., Matsudaira,P. and Darnell,J. (1995) *Molecular Cell Biology, Chap.3*, 3rd edn. Scientific American Books, New York.
Matsuda,S., Vert,J.P., Saigo,H., Ueda,N., Toh,H. and Akutsu,T. (2005) *Protein Sci.*, **14**, 2804–2813.
Mondal,S., Bhavna,R., Mohan Babu,R. and Ramakumar,S. (2006) *J. Theor. Biol.*, **243**, 252–260.
Murphy,R.F., Boland,M.V. and Velliste,M. (2000) *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **8**, 251–259.
Nair,R. and Rost,B. (2002) *Protein Sci.*, **11**, 2836–2847.
Nakai,K. (2000) *Adv. Protein Chem.*, **54**, 277–344.
Nakai,K. and Horton,P. (1999) *Trends Biochem. Sci.*, **24**, 34–36.
Nakai,K. and Kanehisa,M. (1992) *Genomics*, **14**, 897–911.
Nakashima,H. and Nishikawa,K. (1994) *J. Mol. Biol.*, **238**, 54–61.
Nakashima,H., Nishikawa,K. and Ooi,T. (1986) *J. Biochem.*, **99**, 152–162.
Pan,Y.X., Zhang,Z.Z., Guo,Z.M., Feng,G.Y., Huang,Z.D. and He,L. (2003) *J.Protein Chem.*, **22**, 395–402.
Park,K.J. and Kanehisa,M. (2003) *Bioinformatics*, **19**, 1656–1663.
Pierleoni,A., Martelli,P.L., Fariselli,P. and Casadio,R. (2006) *Bioinformatics*, **22**, e408–e416.
Pu,X., Guo,J., Leung,H. and Lin,Y. (2007) *J. Theor. Biol.*, **247**, 259–265.
Schaffer,A.A., Aravind,L., Madden,T.L., Shavirin,S., Spouge,J.L., Wolf,Y.I., Koonin,E.V. and Altschul,S.F. (2001) *Nucleic Acids Res.*, **29**, 2994–3005.
Shen,H.B. and Chou,K.C. (2005a) *Biochem. Biophys. Res. Comm.*, **337**, 752–756.

Shen,H.B. and Chou,K.C. (2005b) *Biochem. Biophys. Res. Commun.*, **334**, 288–292.
Shen,H.B. and Chou,K.C. (2007) *Biochem. Biophys. Res. Commun.*, **355**, 1006–1011.
Shi,J.Y., Zhang,S.W., Pan,Q., Cheng,Y.-M. and Xie,J. (2007) *Amino Acids*, doi 10.1007/s00726-006-0475-y.
Spector,D.L. (2001) *J. Cell. Sci.*, **114**, 2891–2893.
Vapnik,V. (1998) *Statistical Learning Theory*. Wiley-Interscience, New York.
Xiao,X., Shao,S., Ding,Y., Huang,Z., Huang,Y. and Chou,K.C. (2005) *Amino Acids*, **28**, 57–61.
Yuan,Z. (1999) *FEBS Lett.*, **451**, 23–26.
Zhang,S.W., Pan,Q., Zhang,H.C., Shao,Z.C. and Shi,J.Y. (2006a) *Amino Acids*, **30**, 461–468.
Zhang,T., Ding,Y. and Chou,K.C. (2006b) *Comput. Biol. Chem.*, **30**, 367–371.
Zhang,Z.H., Wang,Z.H., Zhang,Z.R. and Wang,Y.X. (2006c) *FEBS Lett.*, **580**, 6169–6174.
Zhou,G.P. (1998) *J. Protein Chem.*, **17**, 729–738.
Zhou,G.P. and Doctor,K. (2003) *Proteins Struct. Funct. Genet.*, **50**, 44–48.

Received July 16, 2007; revised August 13, 2007; accepted September 13, 2007

Edited by Peter Hudson