

# Prediction of Protein Nuclear Localization Based on Smoothed Evolutionary Information and Probabilistic Latent Semantic Indexing

Emily Chia-Yu Su<sup>1</sup>  
cysu@iis.sinica.edu.tw

Jia-Ming Chang<sup>2</sup>  
chang.jiaming@gmail.com

Ting-Yi Sung<sup>1</sup>  
tsung@iis.sinica.edu.tw

Wen-Lian Hsu<sup>1</sup>  
hsu@iis.sinica.edu.tw

<sup>1</sup> Bioinformatics Lab., Institute of Information Science, Academia Sinica, Taipei, Taiwan

<sup>2</sup> Comparative Bioinformatics Group, Bioinformatics and Genomics Programme, Centre for Genomic Regulation, Barcelona, Spain

## Abstract

Nuclear localization of proteins is crucial to various molecular functions and biological processes within a eukaryotic cell. We present a nuclear localization prediction method, in which smoothed evolutionary information and probabilistic latent semantic indexing are incorporated. Experiment results show that our method significantly enhances the prediction performance by 0.051 and 0.098 in terms of overall accuracy and Matthew's correlation coefficient, respectively. Moreover, the proposed biological features are interpretable and can be applied for further experimental studies of nuclear targeting signals.

**Keywords:** protein subcellular localization prediction, nuclear localization, probabilistic latent semantic indexing, smoothed position specific score matrix, support vector machines

## 1 Introduction

The study of protein subcellular localization is important for elucidating protein functions and interactions involved in various cellular processes. However, determination of subcellular localization using experimental approaches is time-consuming. Thus, using computational approaches to predict localization efficiently has become highly desirable. An important compartment in eukaryotic cells is the nucleus, where proteins containing targeting signals, such as nuclear localization signals (NLS) and nuclear export signals (NES), shuttle between the cytoplasm. At present, only three predictors, PredictNLS [1], NucPred [2], and NUCLEO [3], have been designed to specifically identify nuclear proteins. However, they only obtained overall accuracy and Matthew's correlation coefficient (*MCC*) in the range of 0.54~0.70 and 0.25~0.38 [3], respectively.

## 2 Method and Results

We present a method, PSLNuc, to predict protein nuclear localization based on smoothed evolutionary information [4] and probabilistic latent semantic indexing [5]. PSLNuc extracts biological features from gapped-dipeptides of various distances, where smoothed evolutionary information from the position specific score matrix (PSSM) is utilized to determine the weighting of each gapped-dipeptide. Then, the features are further reduced by probabilistic latent semantic indexing (PLSI) and incorporated as input vectors for support vector machine classifiers. The standard PSSM assumes independency between different residues, however, this simplifying assumption does not hold. It has been showed that smoothed PSSM encoding scheme can significantly enhance the performance of RNA-binding site prediction in proteins [4]. In the prediction of nuclear localization,

we also incorporate smoothed PSSM encoding to consider the correlation and dependency from the neighboring residues for each amino acid in a protein.

Experiment results as illustrated in Table 1 show that PSLNuc significantly enhances the prediction performance. Using five-fold cross-validation on a non-redundant (<10% sequence identity) benchmark data set, PSLNuc achieves 0.800 and 0.595 in overall accuracy and *MCC*, respectively, compared favorably to the state-of-the-art results of 0.749 and 0.497. Evaluated by an independent test set, our method also performs better than other approaches by 0.039~0.199 and 0.077~0.207 in terms of overall accuracy and *MCC*, respectively. Our results lend support to the assumption that smoothed PSSM encoding can better resolve the ambiguity of discriminating between nuclear and non-nuclear proteins by considering the dependency from surrounding residues. Moreover, the proposed biological features and gapped-dipeptide signatures are interpretable and can be applied in experimental studies of targeting signals, including NLS and NES.

Table 1: Performance comparison of nuclear localization predictors.

Method	Training Set (by five-fold cross-validation)							
	tp	tn	fp	fn	Sens	Spec	Acc	<i>MCC</i>
PSLNuc	2317	2030	576	525	<b>0.815</b>	<b>0.779</b>	<b>0.800</b>	<b>0.595</b>
NUCLEO	2157	1924	682	685	0.759	0.760	0.749	0.497
Method	Independent Text Set							
	tp	tn	fp	fn	Sens	Spec	Acc	<i>MCC</i>
PSLNuc	452	258	140	112	<b>0.805</b>	<b>0.646</b>	<b>0.739</b>	<b>0.457</b>
NUCLEO	430	246	152	134	0.760	0.620	0.700	0.380
PredictNLS	153	369	29	411	0.270	0.930	0.540	0.250
NucPred	376	233	165	188	0.670	0.590	0.630	0.250

## References

- [1] Cokol, M., Nair, R., and Rost, B., Finding nuclear localization signals, *EMBO Rep.*, 1(5):411, 2000.
- [2] Brameier, M., Krings, A., and MacCallum, R.M., NucPred-Predicting nuclear localization of proteins, *Bioinformatics*, 23(9):1159, 2007.
- [3] Hawkins, J., Davis, L., and Boden, M., Predicting nuclear localization, *J. Proteome Res.*, 6(4):1402, 2007.
- [4] Cheng, C.W., Su, E.C.Y., Sung, T.Y., and Hsu, W.L., Predicting RNA-binding sites of proteins using support vector machines and evolutionary information, to appear in *BMC Bioinformatics*, 2008.
- [5] Chang, J.M., Su, E.C.Y., Lo, A., Chiu, H.S., Sung, T.Y., and Hsu, W.L., PSLDoc: Protein subcellular localization prediction based on gapped-dipeptides and probabilistic latent semantic analysis, *PROTEINS: Structure, Function, and Bioinformatics*, 72(2):693, 2008.