NIM, A Novel Computational Method for Predicting Nuclear-Encoded Chloroplast Proteins

Jun Ding and Haiyan Hu
Department of Electrical Engineering and Computer Science, University of Central Florida, Orlando, USA
Email: {jding, haihu} @cs.ucf.edu

Xiaoman Li[1]
Burnett School of Biomedical Science, University of Central Florida, Orlando, USA
Email: xiaoman@mail.ucf.edu

***Abstract*—The identification of nuclear-encoded chloroplast proteins is important for the understanding of their functions and their interaction in chloroplasts. Despite various endeavors in predicting these proteins, there is still room for developing novel computational methods for further improving the prediction accuracy. Here we developed a novel computational method called NIM based on interpolated Markov chains to predict nuclear-encoded chloroplast proteins. By testing the method on real data, we show NIM has an average sensitivity larger than 92% and an average specificity larger than 97%. Compared with the state-of-the-art methods, we demonstrate that NIM performs better or is at least comparable with them. Our study thus provides a novel and useful tool for the prediction of nuclear-encoded chloroplast proteins.**

***Index Terms*—Nuclear-encoded chloroplast proteins, interpolated Markov chains, subcellular localization, maize, rice**

I. Introduction

The prediction of nuclear-encoded chloroplast (NeC) proteins is important for the annotation of plant proteins and for the study of chloroplast functions [1]. NeC proteins are proteins encoded by genes in the nuclear genome and perform their functions in the chloroplasts of plant cells (Fig. 1) [2]. MRNAs are transcribed from the genes encoding NeC proteins and become mature mRNAs in cytoplasm, which are then translated into proteins in cytoplasm. These NeC proteins are then transferred into chloroplasts from the cytoplasm (Fig. 1). NeC proteins are essential for chloroplast functions, such as photosynthesis and the production of diverse metabolites [3]. Although it is estimated that there are about 3,000 NeC proteins in a land plant species, the majority of NeC proteins are unknown in most of the plant species [2]. For instance, only 54 NeC proteins are annotated in maize in the Uniprot database currently. With the important roles of NeC proteins and the currently limited knowledge of NeC proteins, it is necessary to develop methods to discover NeC proteins in land plant species.

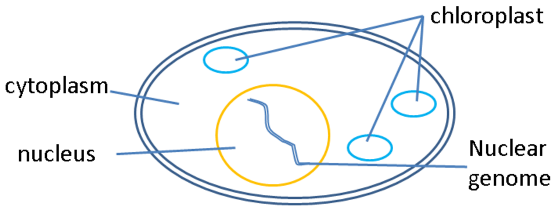

Figure 1. A typical plant cell. The NeC proteins are coded by genes in the nuclear genome of the plant cell and transferred from cytoplasm to chloroplasts after being translated in the cytoplasm of the cell.

Several experimental approaches can be applied to discover NeC proteins, such as tandem mass spectrometry experiments and visualization of fluorescent proteins [4]. In fact, these methods have successfully identified a couple of thousand putative NeC proteins in *Arabidopsis* and enable unprecedented opportunity to study NeC proteins in *Arabidopsis* [2], [4]. However, these experimental techniques are still costly and noisy [4]. It is thus necessary to develop computational methods to predict NeC proteins.

A plethora of computational methods [5]-[11] have been developed to predict subcellular localization of proteins, including chloroplasts. Based on the types of information used for making the prediction, these methods can be classified into three types. The first type is the sorting signal based methods. The sorting signal is the sequence patterns in the N-terminal or C-terminal sequence of a protein that tells the protein to go to the proper subcellular locations in a cell. For instance, the sorting signal of NeC proteins are in the N-terminal sequence of the Nec proteins, because the N-terminal sequences of the majority of NeC proteins are necessary and sufficient for the NeC proteins to enter chloroplasts from cytoplasm [12]. These N-terminal sequences of NeC proteins are called chloroplast transit peptide (cTPs). The discovery of sequence patterns in cTPs could thus enable the identification of NeC proteins. The second type is the amino acid composition based methods. These methods utilize the frequency of individual amino acids or dipeptides (pairs of consecutive amino acids in sequences) to distinguish NeC proteins from other proteins. The third type of methods is the homology based methods, which

Manuscript received January 12, 2013; revised March 28, 2013.

doi: 10.12720/jomb.2.2.115-119

utilize the information of homologs with annotated localization to predict the localization of a protein.

Despite the existence of methods for NeC protein prediction, more accurate computational methods are needed, due to the following reasons. First, the sorting signal based methods rarely achieve a true positive rate over 80% while simultaneously have a false positive rate of less than 10%, as shown by previous studies [13], [14]. Second, the prediction accuracy of the amino acid composition based methods is usually under 80% [14], [15], due to the fact that sequence patterns other than amino acid composition are also important for the transition of NeC proteins from cytoplasm to chloroplasts [16]. Third, the accuracy of the homology based methods depends on the availability of annotated homologs. Such homology information is often unavailable for NeC proteins in sequenced plant species, which prevents the accurate prediction based on annotated homologs.

Here, we propose a novel computational method called NIM that is based on Interpolated Markov Chains (IMCs) [17]. NIM stands for “NeC protein prediction using IMC Models”. It combines the strategy of the amino acid composition based methods and that of the sorting signal based methods. By considering peptide segments of different lengths in cTPs using IMCs, we are able to effectively extract features in the N-terminal of NeC proteins. By testing NIM on the protein sequences from both the maize genome and the rice genome, we show that it accurately predicts NeC proteins. Compared with several state-of-the-art methods, we demonstrate that NIM works better or is at least comparable with available methods on these datasets. Our study thus provides an efficient method for predicting NeC proteins in sequenced plant species, in which proteins are largely un-annotated. The NIM software tool and the datasets we used can be downloaded from http://www.cs.ucf.edu/~xiaoman/NIM/.

II. Matrials and Methods

A. Data

We obtained all maize and rice NeC proteins with experimentally verified cTP information from the Uniprot Database (http://www.uniprot.org/). All obtained NeC proteins are annotated with direct experimental evidence. In total, we obtained 54 maize NeC proteins and 205 rice NeC proteins. For NeC proteins in maize, we randomly chose 30 of them for training, and used the remaining 24 for testing. We obtained 5 groups of training and testing datasets. Similarly, we obtained 5 groups of training and testing datasets for NeC proteins in rice by randomly choosing 100 of them for training and using the remaining 105 for testing.

In addition to the above training and testing data, we also obtained additional testing data as follows. Assuming that proteins whose annotated cellular locations are not chloroplasts are unlikely present in chloroplasts, we collected all such proteins in maize and rice. We then randomly chose 200 of them as non-NeC proteins in both maize and rice.

B. Markov Chains and IMCs (Interpolated Markov Chains)

A Markov chain is a statistical system, in which the next state of the system only depends on a fixed number of past states. This fixed number is called the order of the Markov chain and the probabilities transiting from a fixed number of past states to the next state are called transition probabilities. In general, a higher-order Markov chain is always better than a lower-order one to predict which state will occur next (i.e., which amino acid will occur next in a cTP sequence). This is because if the states are generated by a lower-order chain, both the higher-order and the lower-order chain can predict the next state in exactly the same way. If the states are generated by a higher-order chain, the lower-order chain cannot estimate the transition probabilities accurately while the higher-order chain can. In practice, however, to infer the transition probabilities in a higher-order Markov chain, one always needs a much larger number of training data than to infer those in a lower-order Markov chain. If the amount of the training data is limited, only some transition probabilities in the higher-order Markov chains can be estimated accurately and one thus needs to use lower-order Markov chains if a fixed-order Markov chain is used. In this way, the next states may not be effectively inferred.

To effectively infer the next state with limited training data, which is the case here to predict NeC proteins with limited cTP sequences, we propose to develop a novel method based on a IMC instead of a fixed-order Markov chain. An IMC combines different orders of Markov chains to model the occurrence of a series of events, such as the occurrence of amino acids in a cTP sequence. Intuitively, with limited known cTP sequences, we could not obtain accurate transition probabilities for all state transitions in a high-order chain. This is because these transition probabilities are calculated based on the frequencies of long peptide segments in cTP sequences. The larger the frequencies are, the more accurate the estimated probabilities are. With limited cTP sequences, only a few long peptide segments occur enough times for the accurate estimation of the corresponding transition probabilities in the high-order chain. An IMC thus takes advantage of the frequently occurring long peptide segments, and uses the high-order chain to predict next state transiting from the past states represented by these frequently occurring long peptide segments. When some long peptide segments rarely occur in the training data, an IMC uses a low-order Markov chain to predict the next state from shorter peptide segments. In this way, to predict which amino acid to occur next in a cTP, an IMC combines high- and low-order chains and weight the predictions from different chains based on the frequencies of the past states in the training data.

In detail, in an *m*-order IMC, the probability that an amino acid x_i occurs after the amino acids $x_1\ x_2\ x_3 \ldots x_{i-1}$, denoted as $IMC_m(x_i \mid x_1 x_2 \ldots x_{i-1})$, depends only on the occurrence of $x_{i-m}\ x_{i-4}\ x_{i-3}\ x_{i-2} x_{i-1}.$ This probability can be described by the weighted sum of the probabilities that x_i

is generated by the first, second, ……, and *m*-th order Markov chains according to the following formula:

$$\begin{aligned} &IMC_m(x_i \mid x_1 x_2 \ldots x_{i-1}) \\ &= IMC_m(x_i \mid x_{i-m} x_{i-m+1} \ldots x_{i-1}) \\ &= w(x_{i-m} x_{i-m+1} \ldots x_{i-1}) \Pr_m(x_i \mid x_{i-m} x_{i-m+1} \ldots x_i) \\ &+ [1 - w(x_{i-m} x_{i-m+1} \ldots x_{i-1})] IMC_{m-1}(x_i \mid x_{i-m+1} x_{i-m+1} \ldots x_{i-1}) \end{aligned}$$

In the above formula, the item $w(x_{i-m} x_{i-m+1} \ldots x_1)$ is the weight for using the *m*-order Markov chain to calculate the transition probability from the state specified by $x_{i-m} x_{i-m+1} \ldots x_{i-1}$ to the state specified by x_i. The item $\Pr_m(x_i \mid x_{i-m} x_{i-m+1} \ldots x_i)$ is the transition probability from the m-order Markov chain.

C. Model cTPs by IMCs

To accurately predict NeC proteins, we hope to effectively model how cTPs in NeC proteins are generated. The rationale is that cTPs are necessary and sufficient for NeC proteins to enter chloroplasts [12]. Therefore, if one could model the generation of cTPs, one could predict NeC proteins accurately.

To model cTPs, we calculate the weight and the transition probabilities in the above formula as follows. We count the occurrence number of all *m*-mers, where an *m*-mer is an *m* amino acid long peptide segment in the training cTP sequences. If an *m*-mer occurs more than α times in the training data, its weight will be 1 and the transition probability from this *m*-mer to any amino acid will be estimated from the training data using the m-order Markov chain. Otherwise, the transition probabilities will be calculated similarly, while the weight will be calculated based on chi-square test with the given parameter α, as in [17]. In this case, the weight will be smaller than one and Markov chains with different orders are used to estimate the probability of the next state.

To determine α, we defined the score of each training cTP sequence $x_1 x_2 \ldots x_n$ as $\frac{\sum_{i=m+1}^{n} \ln(IMC_m(x_i \mid x_{i-m} x_{i-m+1} \ldots x_{i-1}))}{n-m+1}$. In this way, the length of a sequence has been taken into account in calculating its score. We then calculate the standard deviations of these scores. We select α such that the scores of the training cTP sequences have the smallest standard deviation. The rationale is that cTPs in different NeC proteins should be similar and thus have similar scores.

D. Predict NeC Proteins

With the weights and transition probabilities calculated from the training data, we predict NeC proteins as follows. Given a protein, we obtain its N-terminal sequence of x amino acid long, where x is the average length of cTPs in the training data. We then calculate the score of this N-terminal sequence based on the weights of different transition probabilities. If the score is larger than a score cutoff, we predict this protein a Nec protein. Otherwise, we consider it as a non-Nec protein.

III. Results

E. Results on Maize Proteins

We tested NIM on the 24 maize NeC proteins. In five experiments, with 24 maize NeC proteins as our testing data, NIM identified 22 or more proteins as NeC proteins (Table 1). In two out of the five experiments, NIM predicted all 24 proteins as NeC proteins. On average, NIM has a sensitivity of 95.0%. We also tested NIM on 200 maize non-NeC proteins. Only 6 out of the 200 proteins were predicted as NeC proteins. This represents a specificity of 97%. Note that some non-NeC proteins may be NeC proteins, due to the incompleteness of current gene annotation.

TABLE I. Prediction Results by NIM and Other Methods.

Datasets Method	54 maize NeC proteins	200 maize non-NeC proteins	205 rice NeC proteins	200 rice non-NeC proteins
NIM	Test1 23/24	6/200=3%	Test1 99/105	3/200=1.5%
	Test2 23/24		Test2 96/105	
	Test3 22/24		Test3 99/105	
	Test4 23/24		Test4 99/105	
	Test5 23/24		Test5 97/105	
	Average 95.0%		Average 92.2%	
TargetP	43/54=79.6%	6/200=3%	182/205=88.8%	7/200=3.5%
AtsubP (composition only)	26/54=48.1%	12/200=6%	105/205=51.2%	33/200=16.5%
AtsubP (Hybrid)	54/54=100%	12/200=6%	181/205=88.3%	11/200=5.5%
WegoLoc	54/54=100%	1/200=0.5%	179/205=87.3%	2/200=1%

We next compared NIM with one of the most widely used method, TargetP [9]. Since TargetP used other NeC proteins for training, we tested TargetP using the 54 maize NeC proteins and 200 non-NeC proteins directly. TargetP predicted 43 out of the 54 NeC proteins and 6 out of the 200 non-NeC proteins as NeC proteins (Table I). Thus, NIM has a higher sensitivity and the same specificity as TargetP.

We further compared NIM with two recently developed methods, AtSubP [14] and WegoLoc [18]. AtSubP uses support vector machines to combine different features such as amino acid composition and homology information to predict protein subcellular localization. We tried two different versions of AtSubP. In the first version, AtSubP only used the amino acid composition information, which had a sensitivity of 48.1% and a specificity of 1-6%=94% (Table I). The much lower sensitivity by using the amino acid composition only in AtSubP demonstrates that IMCs are better model to characterize sequence features than the simple frequencies of individual amino acids or dipeptides. In the second version, AtSubP used all types of information (hybrid), including the homology information. AtSubP had a sensitivity of 100% and a specificity of 6%. It is evident that homology information is the reason for the much improved sensitivity. WegoLoc uses the homology information and the gene ontology annotation to predict protein subcellular localization. When tested on the 54 NeC proteins and 200 non-NeC proteins, WegoLoc has a sensitivity of 100% and a specificity of 1-0.5%=99.5% (Table I). Similar to the hybrid AtSubP, the perfect sensitivity by We goLoc here may be mainly due to the known homology information for the testing data here. In fact, we searched the orthologs of the 54 maize proteins in *Arabidopsis* and found that at least 40 (74.1%) of them have been annotated as NeC proteins in *Arabidopsis* only.

F. Results on the Rice Data

We tested NIM on the 105 NeC proteins and 200 non-NeC proteins in rice. In five experiments, in which 105 rice NeC proteins are used for testing, NIM identified 96 or more proteins as NeC proteins (Table I). On average, NIM has a sensitivity of at least 92.2%. We also tested NIM on 200 rice non-NeC proteins. Only 3 out of the 200 non-NeC proteins were predicted as NeC proteins. This represents a specificity of 1-1.5%=98.5%. Both the sensitivity and the specificity are similar to those obtained on maize data.

Similarly, we compared NIM with TargetP, AtSubP, and WegoLoc on the rice data (Table I). Different from the results on the maize data, we found that our method clearly shown higher sensitivity on the rice data compared with the other three methods. Two factors may contribute to the higher sensitivity in rice than in maize. One is that we have more training data in rice than in maize. Therefore, the transition probabilities and the weights in the obtained IMC from the rice data are more accurate, which results in higher true positive rate. The other is that there are relatively fewer rice NeC proteins with orthologous proteins annotated as NeC proteins. For instance, for the 205 rice NeC proteins, only about 108 (52.7%) have orthologs in *Arabidopsis* that are annotated as NeC proteins, which may result in lower sensitivity of AtSubP and WegoLoc (based on the ortholog information in the Ensembl Plants database). Besides higher sensitivity, NIM had a comparable specificity as other three methods (Table I).

IV. Discussion

We developed a novel method called NIM for predicting NeC proteins in plants. By testing it on maize and rice data, we shown that NIM is better than or at least comparable with the state-of-the-art methods. Our study thus provides a useful tool for subcellular localization prediction. The tool and the datasets we used are freely available at http://www.cs.ucf.edu/~xiaoman/NIM/.

In this paper, we only applied NIM to predict NeC proteins. Since the essence of the method is to combine Markov chains with different orders to describe the common features in training sequences, it should be readily applicable to predict proteins in other subcellular localizations. In addition, we only tested NIM in maize and rice. Even with a small number of training data in maize, the performance of NIM is comparable with other methods. NIM will thus be useful for predicting NeC proteins in recently sequenced plant species [19], in which the majority of NeC proteins may have no annotated homologs.

Acknowledgment

This work was supported in part by grants from the United States National Science Foundation (1125676 (HH), 1218275 (HH), and 1218275 (XL)).

References

[1] P. Jarvis, “Targeting of nucleus-encoded proteins to chloroplasts in plants,” *The New Phytologist*, vol. 179, no. 2, pp. 257-285, 2008.

[2] E. Richly and E. Leister, “An improved prediction of chloroplast proteins reveals diversities and commonalities in the chloroplast proteomes of Arabidopsis and rice,” *Gene*, vol. 329, pp. 11-16, 2004.

[3] F. Myouga, K. Akiyama, R. Motohashi, T. Kuromori, and T. Ito, *et al*, “The Chloroplast Function Database: A large-scale collection of Arabidopsis Ds/Spm- or T-DNA-tagged homozygous lines for nuclear-encoded chloroplast proteins, and their systematic phenotype analysis,” *Plant J*, vol. 61, no. 3, pp. 529-542, 2010.

[4] J. L. Heazlewood, R. E. Verboom, J. Tonti-Filippini, I. Small, and A. H. Millar, “SUBA: The arabidopsis subcellular database,” *Nucleic Acids Research*, vol. 35(Database issue), pp. D213-218, 2007.

[5] K. C. Chou and H. B. Shen, “Large-scale plant protein subcellular location prediction,” *Journal of Cellular Biochemistry*, vol. 100, no. 3, pp. 665-678, 2007.

[6] R Nair and B. Rost, “Mimicking cellular sorting improves prediction of subcellular localization,” *Journal of Molecular Biology*, vol. 348, no. 1, pp. 85-100, 2005.

[7] A. Hoglund, P. Donnes, T. Blum, H. W. Adolph, and O. Kohlbacher, “MultiLoc: Prediction of protein subcellular localization using N-terminal targeting sequences, sequence motifs and amino acid composition,” *Bioinformatics (Oxford, England)*, vol. 22, no. 10, pp. 1158-1165, 2006.

[8] P. Horton, K. J. Park, T. Obayashi, N. Fujita, H. Harada, and C. J. Adams-Collier, *et al*, “WoLF PSORT: Protein localization predictor,” *Nucleic Acids Research*, vol. 35, pp. 585-587, 2007.

[9] O. Emanuelsson, S. Brunak, G. Von Heijne, and H. Nielsen, “Locating proteins in the cell using targetp, signalp and related tools,” *Nature Protocols*, vol. 2, no. 4, pp. 953-971, 2007.

[10] Z. Lu, D. Szafron, R. Greiner, P. Lu, D. S. Wishart, and B. Poulin, *et al*, “Predicting subcellular localization of proteins using machine-learned classifiers,” *Bioinformatics (Oxford, England)*, vol. 20, no. 4, pp. 547-556, 2004.

[11] I. Small, N. Peeters, F. Legeai, and C. Lurin, “Predotar: A tool for rapidly screening proteomes for N-terminal targeting sequences,” *Proteomics*, vol. 4, no. 6, pp. 1581-1590, 2004.

[12] B. D. Bruce, “Chloroplast transit peptides: structure, function and evolution,” *Trends in Cell Biology*, vol. 10, no. 10, pp. 440-447, 2000.

[13] R. Mott, J. Schultz, P. Bork, and C. P. Ponting, “Predicting protein cellular localization using a domain projection method,” *Genome Research*, vol: 12(8), pp: 1168-1174, 2002.

[14] R. Kaundal, R. Saini, and P. X. Zhao, “Combining machine learning and homology-based approaches to accurately predict subcellular localization in Arabidopsis,” *Plant Physiology*, vol. 154, no. 1, pp. 36-54, 2010.

[15] S. Hua and Z. Sun, “Support vector machine approach for protein subcellular localization prediction,” *Bioinformatics (Oxford, England)*, vol. 17, no. 8, pp. 721-728, 2001.

[16] D. W. Lee, S. Lee, G. J. Lee, K. H. Lee, S. Kim, and G. W. Cheong, *et al*, “Functional characterization of sequence motifs in the transit peptide of Arabidopsis small subunit of rubisco,” *Plant Physiology*, vol. 140, no. 2, pp. 466-483, 2006.

[17] S. L. Salzberg, A. L. Delcher, S. Kasif, and O. White, “Microbial gene identification using interpolated Markov models,” *Nucleic acids Research*, vol. 26, no. 2, pp. 544-548, 1998.

[18] S. M. Chi and D. Nam, “WegoLoc: Accurate prediction of protein subcellular localization using weighted gene ontology terms,” *Bioinformatics (Oxford, England)*, vol. 28, no. 7, pp. 1028-1030, 2012.

[19] D. M. Goodstein, S. Shu, R. Howson, R. Neupane, R. D. Hayes, and J. Fazo, *et al*, “Phytozome: a comparative platform for green plant genomics,” *Nucleic acids Research*, vol. 40, pp. 1178-1186, 2012.

Ding Jun, a third-year PhD candidate student at Department of Elec trical Engineering and Computer Science, University of Central Florida. He is majorly working on computational biology and bioinformatics.

Haiyan Hu is an Assistant Professor at Department of Electrical Engineering and Computer Science, University of Central Florida. She is working on bioinformatics and data mining.

Xiaoman Li is an Assistant Professor at Burnett School of Biomedical Science, University of Central Florida. He is working on bioinformatics and statistics.