Prediction of Nuclear Localization Signals by HMM

Keun-Joon Park park@kuicr.kyoto-u.ac.jp
Minoru Kanehisa kanehisa@kuicr.kyoto-u.ac.jp
Institute for Chemical Research, Kyoto University
Gokasho, Uji, Kyoto 611-0011, Japan

1 Introduction

The transport of nuclear proteins is a very important step for the maintenance of life. Most proteins that function in the nucleus are moved through the nuclear envelope that is penetrated by the nuclear pores. The nuclear pore is formed by a large, complex structure known as the nuclear pore complex (NPC). The NPC is a major gateway for mediating ions, small molecules, proteins, RNAs, and ribonucleoprotein particles in and out of the nucleus.

The nuclear proteins are transported to the nucleus via the NPC, if they contain nuclear localization signals (NLSs). NLS was first found in SV40 T antigen as a short cluster of five contiguous positively charged residues in the sequence $_{126}$Pro-Lys-Lys-Lys-Arg-Lys-Val$_{132}$. A family of simple NLSs of this type was generally characterized by one short basic stretch of sequence (4-8 residues) containing several lysine and arginine residues. Another typical NLS, known as bipartite NLS motif contains two interdependent positively charged clusters separated by a mutation tolerant linker region of 10-12 amino acids (Table 1). Both variants of the NLS are recognized by importin alpha family. The precise locations of NLSs within the amino acid sequence of the nuclear proteins are not regular unlike other localization signal peptides.

Table 1: Simple and Bipartite Nuclear Localization Signals

NLS Source	**NLS**
Simple NLSs	
SV40	PKKKRKV (D. Kalderon *et al.*, 1984)
H2B	GKKRSKV (R.B. Moreland *et al.*, 1987)
v-Jun	KSRKRKL (K. Chida and P. K. Vogt., 1992)
Bipartite NLSs	
Nucleoplasmin	KRPAATKKAGQAKKKKLDK (J. Robbins *et al.*, 1991)
NIN2	RKKRKTEEESPLKDKAKKSK (J.A. Kleinschmidt and A. Seiter, 1988)
SWI5	KKYENVVIKRSPRKRGRPRK (D.A. Janes *et al.*, 1995)

In this paper we describe a method for predicting the location of NLS. For this purpose, we constructed the training data set from the SWISS-PROT protein sequence database release 38. This data set consists of 273 simple NLS entries and 87 bipartite NLS entries.

2 Method

In building the data set, we did not use proteins that were not entered in SWISS-PROT database because we wanted to know the whole sequence of the proteins. We first made the list of proteins that contained NLS both from the literature and the SWISS-PROT database. For protein sequences with

the same gene but from different species, only one of them was included. The location of the NLSs and the amino acid sequences of the NLSs were added to the data set. We also entered references on NLSs.

For the NLS prediction we used the Hidden Markov Model (HMM), and the HMM program, HMMER version 2.0 for protein and nucleic acid sequence analysis. HMMER is freely available from http://hmmer.wustl.edu/. The HMMs were trained on the unaligned sequences of NLSs in our data set.

In order to distinguish NLSs, positively charged residues Arg (R) and Lys (K) may be considered to have the same function. In this work, we have extended this notion by grouping of amino acids in the sequences of the data sets and query proteins. The performance of the HMMs to recognize NLSs trained on the data sets was improved by reducing the number of different kinds of amino acids (alphabet characters). For example, we found that the grouping of [KR], [DE], P, M, F, and X (the rest) was one of the most effective combinations. We could obtain the result of NLS prediction as the score which is related to the statistical significance of the alignment.

3 Result and Discussion

To test the performance of the HMM, protein sequences in the data set were divided into two sub-data sets, which were used as the training and testing data sets. The rate of correct prediction of original 20 residue HMM without grouping was only 13.9%. At the same conditions, another HMM with a grouping of [KR], [DE], P, M, F, Xs showed 89.4% prediction accuracy.

Due to the difficulties in the experimental determination of protein's cellular localization, the methods of theoretical prediction on the known sequence are becoming more important [1]. In some nuclear protein sequences, many possible NLSs are obtained. But in many cases, there is no experimental data provided to us to predict the NLS. Therefore, we cannot tell which is better when faced with many possible NLSs derived from the same nuclear protein sequence. However, for a given protein sequence, we can give the most probable solution based on our method.

Acknowledgments

This work was supported in part by a Grant-in-Aid for Scientific Research on Priority Areas, 'Genome Science', from the Ministry of Education, Science, Sports and Culture of Japan. The computation time was provided by the Supercomputer Laboratory, Institute for Chemical Research, Kyoto University.

References

[1] Nakai, K. and Horton, P., PSORT: a program for detecting the sorting signals of proteins and predicting their subcellular localization, *Trends Biochem. Sci.*, 24(1):34–35, 1999.