1-1-2013

Computational Analysis and Prediction of Genome-Wide Protein Targeting Signals and Localization

JHIH-RONG LIN
University of South Carolina - Columbia

Computational Analysis and Prediction of Genome-Wide Protein Targeting Signals and Localization

by

Jhih-Rong Lin

Bachelor of Engineering
National Tsing Hua University, 2000

Master of Engineering
National Tsing Hua University, 2003

Master of Science
California State University, East Bay, 2009

Submitted in Partial Fulfillment of the Requirements

For the Degree of Doctor of Philosophy in

Computer Science and Engineering

College of Engineering and Computing

University of South Carolina

2013

Accepted by:

Jianjun Hu, Major Processor

John Rose, Committee Member

Homayoun Valafar, Committee Member

Jijun Tang, Committee Member

Hongmei Zhang, Committee Member

Lacy Ford, Vice Provost and Dean of Graduate Studies

Abstract

Computational prediction of protein subcellular localization can greatly help to elucidate its functions. Despite the existence of dozens of protein localization prediction algorithms, the prediction accuracy and coverage are still low. Several ensemble algorithms have been proposed to improve the prediction performance, which usually include as many as 10 or more individual localization algorithms. However, their performance is still limited by the running complexity and redundancy among individual prediction algorithms. In the first part of the dissertation, we propose a novel method for rational design of minimalist ensemble algorithms for practical genome-wide protein subcellular localization prediction. The algorithm is based on combining a feature selection based filter and a logistic regression classifier. Using a novel concept of contribution scores, we analyzed issues of algorithm redundancy, consensus mistakes, and algorithm complementarity in designing ensemble algorithms. We applied the proposed minimalist logistic regression (LR) ensemble algorithm to two genome-wide datasets of Yeast and Human and compared its performance with current ensemble algorithms. Experimental results showed that the minimalist ensemble algorithm can achieve high prediction accuracy with only 1/3 to 1/2 of individual predictors of current ensemble algorithms, which greatly reduces computational complexity and running time. Compared to the best individual predictor, our ensemble algorithm improved the prediction accuracy from AUC score of 0.558 to 0.707 for the Yeast dataset and from 0.628 to 0.646 for the Human dataset.

In the second part of the dissertation, we propose a computational method, SeqNLS,

to predict nuclear localization signal (NLS). The major difficulty of NLS prediction is that NLSs are known to have diverse patterns, but the knowledge to NLS patterns is limited and only a portion of NLSs can be covered by the known NLS motifs. In SeqNLS, on the one hand we propose a sequential-pattern approach to effectively detect potential NLS segments without constrained by the limited knowledge of NLS patterns. On the other hand, we introduce a model for NLS prediction which utilizes the fact that NLS is one type of linear motifs. Our experiment results show that our sequential-pattern approach is effectively in extensively searching potential NLSs. Our method can consistently find over 50% of NLSs with prediction precision at least 0.7 in the two independent datasets. The performance of our method can out perform the-state-of-art NLS prediction methods in terms of F1-score.

The binding affinity between a nuclear localization signal (NLS) and its import receptor is closely related to corresponding nuclear import activity. PTM based modulation of the NLS binding affinity to the import receptor is one of the most understood mechanisms to regulate nuclear import of proteins. However, identification of such regulation mechanisms is challenging due to the difficulty of assessing the impact of the PTM on corresponding nuclear import activities. In the third part of the dissertation we proposed NIpredict, an effective algorithm to predict nuclear import activity given its NLS, in which molecular interaction energy components (MIECs) were used to characterize the NLS-import receptor interaction, and the support vector regression machine (SVR) was used to learn the relationship between the characterized NLS-import receptor interaction and the corresponding nuclear import activity. Our experiments showed that nuclear import activity change due to NLS change could be accurately predicted by the NIpredict algorithm. Based on NIpredict, we developed a systematic framework to identify potential PTM-based nuclear import regulations for human and yeast nuclear proteins. Application of this approach has uncovered the potential nuclear import regulation mechanisms by phosphorylation and/or acetylation

of three nuclear proteins including SF1, histone H1, and ORC6.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background

A eukaryotic cell is organized into different membrane surrounded compartments containing characteristic proteins and performing specialized functions. Functions of proteins are thus closely associated with their subcellular locations. With enormous amount of sequences emerged from the genome sequencing projects, it becomes increasingly important to develop practical tools for functional annotation based on the relevant features from sequences such as localization. Although experimental high-throughput approaches have been developed and applied to determine proteins localization [1, 2], they are extremely expensive and time consuming. Fast, accurate and genome-scale computational methods for predicting subcellular localization of proteins provide an attractive complement to experimental methods.

On the other hand, most proteins are synthesized in the cytoplasm and are transported to their target subcellular locations. The translocations of nascent proteins are usually guided by targeting signals encoded within the amino acid sequences of proteins. Genome-wide identification and deciphering of those targeting signals are important for inferring localization of proteins and understanding the transporting mechanism. Experimentally identifying protein targeting signals is usually done by mutating a target segment in the sequence and then checking its effect on the delivered location [3, 4], which is extremely labor-intensive and expensive; computational

prediction of targeting signals provides a complementary method to assist biologists to design the experiments.

Computational prediction of sorting signals has substantially reduced time and cost for biologists to discover sorting signals by focusing their experiments on putative motifs. However, protein localization is more than an issue of binary outcomes (either localized or not localized to a target compartment). Localizations of certain proteins are regulated through interactions of their sorting signals with other proteins during specific cell cycle(s) [5,6]. Deregulation of such machanisms is associated with numerous cancers such as breast cancers, prostate cancers, and other diseases [7–10].

1.2 Motivation

In the past ten years, dozens of protein localization algorithms have been proposed based on different information sources such as amino acid composition, sorting signals, functional motifs, conserved domains, homology search, and protein-protein interaction [11]. A variety of machine learning techniques, such as SVM and K-nearest neighbour classifiers, have been used in these prediction algorithms. Although existent methods have achieved success at different degrees, a comprehensive evaluation study has shown that many of the reported prediction accuracies are far from being sufficient for genome wide protein localization prediction [12]. Recently, several research groups proposed to apply ensemble or integration of algorithms to protein localization prediction [13–17]. Different ensemble algorithms are used in those studies such as weighted and adaptive weighted voting [13], protocol-based ensemble algorithm [14], Linear Discriminant Analysis (LDA) [15], J48 decision tree (DT) [16], and two-layer decision tree [17]. Most of these ensemble algorithms integrated 10 or more standalone prediction methods for localization prediction without considering their relationships such as redundancy and complementarity. This makes these ensemble algorithms computationally intensive. Furthermore, incorporation of unnec-

essary predictors into an ensemble algorithm may overfit the training data and result in degradation of its prediction performance, which has been reported recently for ensemble mitochondrion predictors [18]. in the first part of the dissertation, we proposed a systematic work to evaluate 9 standalone localization prediction algorithms and analyze their bias and relationships in the prediction space of the resulting ensemble algorithms. We found that ensemble algorithms based on the combination of several specific predictors achieved comparable prediction performance as using all 9 predictors, suggesting that a high degree of redundancy exists among all individual predictors. We thus proposed a minimalist ensemble prediction algorithm for subcellular localization prediction and evaluated its performance on two data sets, which showed high performance and significant reduction of computational complexity and running time.

On the other hand, compared to DNA regulatory motifs, computational prediction of targeting signals remains challenging due to their low conservation at the amino acid level. In the case of Nuclear Localization Signal (NLS), several NLS prediction methods have been developed such as PSORT II [19], PredictNLS [20], NLStradamus [21], cNLS Mapper [22], and NucImport [23]. PSORT II predicts NLSs based on sequence patterns implemented as three simple rules according to the classification of NLSs [24]; the rules are mainly combinations of clusters of basic amino acids K and R and gaps between clusters. PredictNLS predicts NLSs based on 194 potential NLS motifs, which are derived from 114 experimentally verified NLSs by applying a silico mutagenesis approach. Nguyen Ba et. al. [21] found that NLSs tend to have similar residue frequency distributions which are different from that of background residues. Their NLStradamus algorithm detects NLSs by using a simple two-state or four-state HMMs to accommodate the frequency variations. cNLS Mapper estimates classical NLS (cNLS) functionality of a peptide by calculating the sum of the functional contribution of each residue in the peptide according to the

activity-based profiles, which are obtained from systematic amino-acid-replacement analyses in budding yeast. NucImport [23] builds a Bayesian network to predict nuclear localization by incorporating various attributes related to nuclear importing. If a protein is predicted as a nuclear protein, the location of its NLS is predicted as the segment in the protein with the highest cNLS score in the inferred cNLS class based on the dependencies with other attributes in the Bayesian network.

These five NLS prediction methods have achieved success at different degrees. However, their prediction performances are still far from being sufficient to assist biologists to discover putative NLSs in protein sequences of interest. Each of them has their own shortcomings. Although a great portion of NLSs can be covered by the rules used in PSORT II to detect NLS, quite many of the patterns covered by the rules commonly exist in peptides which do not contain NLSs, leading to a high false positive rate. The sensitivity of the PredictNLS algorithm depends on the number of NLS motifs it used, which has been extended by introducing the potential NLS motifs generated using in-silicon mutagenesis analysis. But they are still too specific and couldn't effectively accommodate NLS variability [21]. The performance of the NLStradamus algorithm strongly depends on its assumption that NLSs have certain residue distributions. However, many NLS instances in our testing datasets have shown very different residue frequencies. Both cNLS mapper and NucImport algorithms are developed based on the characteristics of cNLS. However, approximately 43% of all nuclear proteins may use other transport mechanisms other than the classical nuclear import pathway according to Allison Lange et al [25].

All the aforementioned NLS prediction methods heavily rely on sequence features of NLS to predict NLSs. However, NLSs are known to have diverse sequence patterns while the knowledge to NLS sequence patterns is limited. In the second part of the dissertation, we propose a novel algorithm which apply frequent pattern mining techniques to mine sequence patterns within experimentally verified NLSs which can

be used to effectively detect potential NLSs. In addition, we introduce a model which utilizes the fact that NLS is one type of linear motifs. This model can integrate the mined sequence patterns and the linear motif attributes of NLS to effectively predict NLS.

In recent years, an increasing number of researches are devoted to studying nuclear import regulation of proteins. The discovery of the import regulation mechanism for a particular nuclear protein is of great interest since it implies a potential way to control the protein's activity [5]. Moreover, it contributes to uncovering the potential biological pathways that regulate the associated biological activities in the nucleus. Nuclear import activity is mostly regulated through modulating the interactions between nuclear proteins and their binding import receptors [8]. In particular, modulating the NLS binding affinity to its binding receptor by post-translational modification (PTM) is the best understood mechanism (PTM-based nuclear import regulation) that regulates the nuclear import of proteins. In previous studies, the most common type of PTM for nuclear import regulation is phosphorylation [5, 6, 26–28] while lysine acetylation has been found to be another frequent type [29–36]. The reason that nuclear import can be regulated through the PTM is that nuclear import activity is directly related to the binding affinity of NLS for its binding import receptor [37–39]. However, identifying the PTM-based nuclear import regulation is difficult since PTM may promote, repress or may not have obvious impact on the nuclear import activity [27].

The most commonly used approach to infer the PTM-based nuclear import regulation is the site-directed mutagenic analysis [26, 32, 40–46]. This approach basically mutates the NLS residue at the PTM site to a residue that either prevents the PTM or mimics the residue after the PTM. It then evaluates the likelihood that the PTM regulates the nuclear import of the protein based on the change of the corresponding nuclear import activity. The strategy of mimicking residue after PTM such as phosphorylation has been performed computationally by cNLS mapper [22], in which

the position-wise contributions of different amino acids to the nuclear import activity are approximated in the activity-based profiles. However, the interaction between a NLS and its binding import receptor is very sensitive to the NLS change. The site-directed mutagenic analysis is thus not always reliable due to the difference between the mimicking residues and the residues after PTM. Since the PTM-based nuclear import regulation is now recognized as a common nuclear import regulation mechanism, there is a need for developing quantitative methods to expand the identification of more PTM-regulated nuclear proteins [27].

For the PTM-based nuclear import regulation, it is technically true that PTM regulates the nuclear import of a protein through modification of its NLS residue(s). However, the induced change on the interaction between the NLS and the import receptor is the ultimate factor that governs the change on its nuclear import activity. In other words, the induced change on the NLS-import receptor interaction should better characterize the change of the nuclear import activity caused by PTM than the difference of the NLSs. Therefore, in our method we first applied molecular interaction energy components (MIEC) [47–50], which has been successfully used to characterize domain-peptide interactions, to characterize the NLS-import receptor interaction. Next, we used SVR to learn the relationship between the MIEC features and the corresponding nuclear import activity, which is quantitated as NLS activity scores [22] in the experimental dataset. The characteristic of our method (NIpredict) is that it is a machine learning based method based on features calculated from NLS-import receptor interaction interface, which can thus be applied to assess the impact of PTM within NLS on the corresponding nuclear import activity.

1.3 Contributions and Significance

The research presented in this work addresses three major problems discussed in the previous section. In particular, seven major contributions are:

- Analyzing the existing 9 protein localization predictors systematically, which in particular addresses issues of algorithm redundancy, consensus mistakes, and algorithm complementarity in designing ensemble algorithms (in Chapter 2).

- Proposing a novel method for rational design of minimalist ensemble algorithms for practical genome-wide protein subcellular localization prediction, which can significantly reduce the number of individual predictors in a given ensemble algorithm while maintaining comparable performance (in Chapter 2).

- Demonstrating the linear motif attributes of NLS such as disorder, relative surface area, and relatively local conservation (in Chapter 3).

- Proposing an algorithm (SeqNLS) to predict NLS which outperforms other state-of-the-art NLS predictors (in Chapter 3).

- Proposing an algorithm (NIpredict) to predict nuclear import activity effectively based on NLS-import receptor interaction (in Chapter 4).

- Developing a systematic framework to identify potential PTM-based nuclear import regulations for human and yeast nuclear proteins based on NIpredict (in Chapter 4).

- Uncovering the potential nuclear import regulation mechanisms by phosphorylation and/or acetylation of three nuclear proteins including SF1, histone H1, and ORC6 (in Chapter 4).

1.4 Organization of the Dissertation

The rest of the dissertation is organized into five chapters:

Chapter 2 analyzed 9 existing protein localization predictors, which in particular addressed issues of algorithm redundancy, consensus mistakes, and algorithm

complementarity in designing ensemble algorithms. A framework of designing minimalist ensemble algorithms for practical genome-wide protein subcellular localization prediction was proposed, which could significantly reduce the number of individual predictors in a given ensemble algorithm while maintaining comparable performance. The work has been published in BMC Bioinformatics, 2012 [51].

Chapter 3 proposed a NLS prediction algorithm, SeqNLS. The method applied frequent pattern mining techniques to address the issues of diverse patterns of NLS. In addition, we demonstrated the linear motif attributes of NLS and designed an algorithm to incorporate the linear motif features of NLS into our method, which sucessfully improved the NLS prediction accuracy. The work has been accepted in PLoS One, 2013 [52].

Chapter 4 proposed a nuclear import activity prediction algorithm, NIpredict. The prediction is based on characterized NLS-import receptor interaction and can be used to identify nuclear proteins whose nuclear import is regulated by PTM. We applied our method in human and yeast genome and uncovered serveral potential nuclear import regulation mechanisms.

Chapter 5 summarized the main results in the dissertation and presented some conclusions.

Chapter 2

Minimalist Ensemble Algorithms for Genome-wide Protein Localization Prediction

2.1 Background

Functions of proteins are closely correlated with their subcellular locations. For example, Assfalg et al. [53] showed that there exists strong correlation between localization and proteins fold and localization can be utilized to predict structure class of proteins. It is thus desirable to accurately annotate subcellular location of proteins to elucidate their functions. In the past ten years, dozens of protein localization algorithms have been proposed based on different information sources such as amino acid composition, sorting signals, functional motifs, conserved domains, homology search, and protein-protein interaction [11]. A variety of machine learning techniques, such as SVM and K-nearest neighbour classifiers, have been used in these prediction algorithms. Although existent methods have achieved success at different degrees, a comprehensive evaluation study has shown that many of the reported prediction accuracies are far from being sufficient for genome wide protein localization prediction [12].

Recently, several research groups proposed to apply ensemble or integration of algorithms to protein localization prediction [13–17]. Liu et al. [13] proposed weighted and adaptive weighted voting algorithms in which the overall accuracy of a standalone algorithm is used as the weight. Laurila and Vihinen [14] proposed an integrated method (PROlocalizer) which combines the predictions of multiple specialized

binary localization prediction algorithms such as TMHMM and Phobius. Park et al. [15] developed a Linear Discriminant Analysis (LDA) method (ConLoc) to assign LDA optimal weights for weighted voting. Assfalg et al. [16] proposed two ensemble localization algorithms; one is a scored voting scheme based on the ranks of the prediction accuracy of the predictors; the other chose J48 decision tree (DT) classifier as the integration scheme. Shen and Burger [17] proposed a two-layer decision tree method to improve the prediction accuracy of a single subcellular location. Most of these ensemble algorithms integrated 10 or more standalone prediction methods for localization prediction without considering their relationships such as redundancy and complementarity. This makes these ensemble algorithms computationally intensive. Furthermore, incorporation of unnecessary predictors into an ensemble algorithm may overfit the training data and result in degradation of its prediction performance, which has been reported recently for ensemble mitochondrion predictors [18].

In this chapter, we evaluated 9 standalone localization prediction algorithms and analyzed their bias and relationships in the prediction space of the resulting ensemble algorithms. We found that ensemble algorithms based on the combination of several specific predictors achieved comparable prediction performance as using all 9 predictors, suggesting that a high degree of redundancy exists among all individual predictors. We thus proposed a minimalist ensemble prediction algorithm for subcellular localization prediction and evaluated its performance on two data sets, which showed high performance and significant reduction of computational complexity and running time.

2.2 Methods

2.2.1 Standalone protein localization predictors

To implement our ensemble localization predictor, we selected 8 published localization prediction algorithms provided that the software or web server is publicly avail-

Table 2.1: Features used in localization prediction algorithms

	Sorting signal	Amino acid composition	*Known	Homology search	Evolutionary information	PPI
NetLoc						X
YLoc	X	X	X	X		
MultiLoc2	X	X	X		X	
KnowPred				X		
Subcell		X				
WoLFPSORT	X	X	X			
BaCelLo		X			X	
CELLO		X				
SubLoc		X				

*Known domains or motifs

able, and batch submission is supported. These algorithms include YLoc [54], MultiLoc2 [55], KnowPred [56], Subcell [57], WoLFPSORT [58], BaCelLo [59], CELLO [60], SubLoc [61]. We also included NetLoc [62], a protein-protein interaction (PPI) based prediction method. These prediction methods differ in the features that characterize proteins targeting different subcellular locations (Table 2.1) and the prediction algorithms. These diverse features include sorting signals, amino acid composition, known motifs or domains, homology search against a known dataset or database such as SwissProt, evolutionary information such as phylogenetic profiles or sequence profiles, and protein-protein interaction. The overlap of the used features among localization predictors suggests that redundant predictions could be made when these prediction methods are combined to build an ensemble algorithm, which could mislead the prediction behaviour of the resulting ensemble algorithm.

In addition to amino acid sequence information, protein-protein interaction has been known as external information correlated to protein subcellular localization. A number of algorithms have been developed to utilize PPI features to predict protein localization (Hishigaki et al [63], Lee et al [64] and Shin et al [65]). Recently, our group developed NetLoc [62], a kernel-based logistic regression (KLR) method, which can effectively extract PPI features to predict protein localization. Considering that NetLoc simply used PPI as its features, we integrated NetLoc into our ensem-

ble algorithms to compare the ensemble performances with and without a PPI-based predictor. In our experiments, PPI data of NetLoc is based on the whole Saccharomyces cerevisiae physical PPI dataset obtained from BioGRID database [66]. We exclude proteins overlapped with our Yeast datasets from the PPI dataset to ensure independency between the training and testing datasets.

2.2.2 Mapping of subcellular locations

Different localization predictors may have different subcell resolutions. In order to compare their performances on genome wide datasets, we applied a location mapping scheme to map the subcellular locations of standalone predictors to unified 5 locations in the ensemble algorithms, including Cytosol, Mitochondrion, Nucleus, Secretory (secretory pathway), and Others. Six classes of subcellular locations are mapped to Secretory according to [55]: extracellular, plasma membrane, endoplasmic reticulum, golgi apparatus, lysosomal, and vacuolar. Except for Cytosol, Mitochondrion, Nucleus, and Secretory, the remaining subcellular locations are categorized as Others. For example, for CELLO, the following subcellular locations are mapped to Secretory: extra, plas, er, vacu, golgi, and lyso; chlo, pero, and cytos are mapped to Others. For WoLFPSORT, E.R., extr, plas, golg, lyso, and vacu are mapped to Secretory; chlo, cysk, and pero are mapped to Others.

2.2.3 Contribution score

To explore the complementary relationship among the individual predictors used in an ensemble algorithm, we calculated contribution scores [67] of component standalone prediction methods. This measure is used to evaluate the contribution of each individual classifier to the ensemble algorithm, and has been used for pruning large ensemble set. The main idea of the contribution score is that predictors that tend to make correct and minority predictions among other predictors will be scored

higher since they make unique contribution and thus are essential for the ensemble algorithm. On the other hand, predictors with low contribution scores tend to make incorrect and majority predictions. The contribution score of a predictor in an ensemble algorithm is calculated as follows: Contribution score of predictor i =

$$\sum_{j=1}^{N}(\alpha_{ij}(2v_{max}^{(j)} - v_{p_i(\text{protein}_j)}^{(j)}) + \beta_{ij}v_{sec}^{(j)} + \theta_{ij}(v_{correct}^{(j)} - v_{p_i(\text{protein}_j)}^{(j)} - v_{max}^{(j)}))$$

where:

$$\alpha_{ij} = \begin{cases} 1 & \text{if } \mathrm{p_i(protein_j)}=\mathrm{real_j} \text{ and } \mathrm{p_i(protein_j)} \text{ is in the minority group;} \\ 0 & \text{otherwise.} \end{cases}$$

$$\beta_{ij} = \begin{cases} 1 & \text{if } \mathrm{p_i(protein_j)}=\mathrm{real_j} \text{ and } \mathrm{p_i(protein_j)} \text{ is in the majority group;} \\ 0 & \text{otherwise.} \end{cases}$$

$$\theta_{ij} = \begin{cases} 1 & \text{if } \mathrm{p_i(protein_j)} \neq \mathrm{real_j}; \\ 0 & \text{otherwise.} \end{cases}$$

Symbols in the formula are explained as follows: for a protein j, the prediction results of nine predictors in the order of predictor 1 to predictor 9 are Cytosol, Nucleus, Nucleus, Mitochondrion, Nucleus, Cytosol, Nucleus, Nucleus, and Nucleus, while the real localization of protein j is Cytosol. In this case, the majority votes (predictions) are for Nucleus, the number of the majority votes is denoted as $v_{max}^{(j)}$, which is 6; the number of the second majority votes is denoted as $v_{sec}^{(j)}$, which is 2; the number of the correct votes is denoted as $v_{correct}^{(j)}$, which is 2; the prediction result of predictor i is denoted as $\mathrm{p_i(protein_j)}$; the number of predictors having the same prediction result with predictor i is denoted as $v_{p_i(\text{protein}_j)}^{(j)}$. From the formula, we can see that predictor 1 and predictor 6 have the same positive contribution, which is 2*6-2=10; predictor 4 has minor negative contribution, which is -5; predictors 2,3,5,7,8,9 have the most

negative contribution, which is -10. If the dataset used to learn contribution scores has N proteins, then the final contribution score of a predictor is summation of its N contributions. We normalized the final contribution scores (CS) with the formula: $(CS - \mu)/\sigma$ where μ and σ are mean and standard deviation of contribution scores among predictors.

2.2.4 Minimalist ensemble prediction algorithm

Existing ensemble algorithms tend to include as many as possible component classifiers for better prediction performance. However, including redundant predictors not only increases computational complexity and collecting effort, but also may lead to over-fitting [9]. Moreover, predictors with poor performance could mislead the ensemble algorithms especially those using majority voting schemes. It is thus desirable to find the minimal subset of predictors for achieving equally good or better prediction performance. Several strategies can be used to find the minimal set of predictors: exhaustive search of all possible combinations of component predictors, feature selection, and selecting top k most accurate predictors. We did an exhaustive search for all combinations of K individual predictors to build different ensemble algorithms. It shows that combining 6 out of 9 predictors can achieve the best performance when the logistic regression classifier was used to integrate the predictions. However, exhaustive search is a time consuming process especially when the set of available predictors is large. Top-K accuracy selection method is straightforward and fast, but has the limitation of neglecting the redundancy among individual predictors.

Here we proposed a minimalist ensemble design method to approximate the smallest set of predictors with the best possible prediction accuracy. The rationale is to find the smallest subset of predictors whose predictions are highly correlated to the real locations. The minimalist ensemble design problem is similar to feature selection when the prediction labels of individual predictors are considered as features. Here,

we chose the correlation based feature subset evaluator (CfsSubsetEval) [68] as the attribute evaluator to evaluate correlation between a feature subset and the class. Greedy-Stepwise method is used to search optimal feature subsets in different size of K through the space: the starting point of search is set as the set with all available predictors (assume size N). Each time Greedy-Stepwise algorithm will remove one feature or predictor from the set which would produce a reduced set with the highest possible CfsSubsetEval Score. We continue the process until set size is 1, while along the way the predictors in the set with size K are recorded as the output of our minimalist ensemble algorithm. After the K individual predictors are selected based on the training dataset, their predicted localizations for all proteins in the training dataset will be used as features, and a machine learning based classifier, such as naive Bayes, logistic regression, or decision trees is used to train a classifier to predict the final subcellular localization. This method used to select minimalist set of individual predictors can also be used for building ensemble algorithms based on weighted voting or LDA.

2.2.5 Datasets preparation

Two genome-wide protein localization databases are used to build three datasets in our experiments. The yeast dataset is obtained from Huh et al [1]. We excluded proteins localized to Others (after location mapping) and multi-location proteins from the yeast dataset. Two versions of the yeast dataset with different resolutions are prepared; for the low-resolution yeast dataset (Yeast Low-Res), we extracted proteins in Cytosol, Nucleus, Mitochondrion, Secretory after location mapping. For the high-resolution yeast dataset (Yeast High-Res), we extracted proteins in Cytosol, Nucleus, Mitochondrion, ER, Vacuole, Golgi, and Cell Periphery (plasma membrane and extracellular). The Human dataset is obtained from the LOCATE database [69] by extracting proteins in 4 locations (Cytoplasmic, Mitochodria, Nuclear, and Extra-

cellular). Then we removed all multi-location proteins. For both Yeast and Human datasets, Blastclust with 30% sequence identity was used to remove redundant sequences. In addition, proteins overlapped with the training datasets of component predictors in the corresponding ensemble experiment are removed. It should be noted that the Yeast High-Res dataset is highly overlapped with the Yeast Low-Res datasets. The final distribution of proteins in different locations for the three datasets is shown in Table 2.2.

Table 2.2: The distributions of proteins in different locations for the test datasets

Dataset	Cytosol	[4]Mit	Nucleus	Sectory				Total
[1]Yeast	498	175	234	315				1222
Human	361	327	159	458				1305
	Cytosol	[4]Mit	Nucleus	ER	[5]Vac	Golgi	[6]Cell	
[2]Yeast	530	165	233	149	103	33	34	1247
[3]Overlap	451	133	218	132	90	32	0	1056

[1]Yeast-LowRes
[2]Yeast-HighRes
[3]Overlap of Yeast-LowRes and Yeast-HighRes
[4]Mitochondrion
[5]Vacuole
[6]Cell Periphery

2.2.6 Evaluation of individual Predictors and ensemble algorithms

To evaluate the performance of predictors, accuracy and MCC were calculated using the equations below: Accuracy: MCC: where TP, TN, FP, FN means true positive, true negative, false positive and false negative predictions. It should be noted that since localization prediction is a multi-class classification problem, MCC can only be calculated for each location while an overall accuracy can be calculated for each prediction method for a given dataset. In our experiments, 10-fold cross-validation was used to evaluate all the ensemble algorithms.

2.3 Results and Discussion

2.3.1 Evaluation of individual predictors

Table 2.3: Prediction performance(MCC Scores) of individual predictors for the Yeast Low-Res dataset

	Cytosol	Mitochondrion	Nucleus	Secretory	Overall Accuracy
YLoc (2010)	0.146	0.556	0.367	0.314	0.453
NetLoc (2010)	0.270	0.350	0.484	0.473	0.556
MultiLoc2 (2009)	0.268	0.581	0.420	0.339	0.558
KnowPred (2009)	0.286	0.415	0.345	0.534	0.51
Subcell (2008)	0.134	0.243	0.181	0.326	0.399
WoLFPSORT (2007)	0.265	0.549	0.312	0.568	0.484
BaCelLo (2006)	0.164	0.526	0.291	0.339	0.468
CELLO (2006)	0.261	0.547	0.302	0.534	0.493
SubLoc (2001)	0.184	0.354	0.260	0.391	0.439
[1]LR	0.429	0.668	0.476	0.607	0.668
[2]LR	0.504	0.666	0.550	0.664	0.707

[1]LR with 8 predictors without NetLoc
[2]LR with all 9 predictors

We obtained the prediction results on three test datasets (Yeast Low-Res, Yeast High-Res and Human) from the selected individual predictors using the web servers or standalone programs and then evaluated their accuracy and MCC scores. The results of 9 predictors for the Yeast Low-Res dataset are shown in Table 2.3, the results of 6 predictors for the Yeast High-Res dataset are shown in Table 2.4, and the results of 8 predictors for the Human dataset are shown in Table 2.5.

For the Yeast dataset (Table 2.3, 2.4), most algorithms have better performance

Table 2.4: Prediction performance(MCC Scores) of individual predictors for the Yeast High-Res dataset

	[1]Y	[2]M	[3]S	[4]W	[5]C	[6]N	[7]LR	[8]LR
Cytosol	0.441	0.293	0.146	0.251	0.255	0.247	0.459	0.555
*Mito	0.689	0.496	0.251	0.510	0.501	0.318	0.684	0.713
Nucleus	0.405	0.275	0.181	0.311	0.306	0.434	0.351	0.473
ER	0.207	0.203	0.022	0.059	0.000	0.340	0.431	0.463
Vacuole	0.115	0.045	0.034	0.000	0.061	0.189	0.174	0.191
Golgi	0.008	0.010	0.054	0.118	-0.005	0.465	0.038	0.275
Cell Periphery	0.107	0.044	0.068	0.142	0.090	0.449	0.04	0.269
Overall Accuracy	0.506	0.473	0.300	0.362	0.354	0.523	0.585	0.640

*Mitochondrion
[1]YLoc(2010)
[2]MultiLoc2(2009)
[3]Subcell(2008)
[4]WoLFPSORT(2007)
[5]CELLO(2006)
[6]NetLoc(2010)
[7]LR with 5 predictors without NetLoc
[8]LR with all 6 predictors

on predicting Mitochondrion proteins. For the Yeast High-Res dataset (Table 2.4), we can see that all predictors except NetLoc showed poor performance on predicting proteins localized to secretory pathway compartments especially golgi, and cell periphery. This suggests that PPI can be an effective feature for predicting low-resolution compartments. Predictors with relatively high accuracy on the Yeast Low-Res Secretory proteins, such as CELLO and WoLFPSORT, don't have corresponding performance on predicting proteins localized to ER, Golgi, Vacuole in the Yeast High-Res dataset which are highly overlapped with the Yeast Low-Res Secretory proteins (Table 2.3). This means those predictors have difficulties in distinguishing smaller compartments of secretory pathway. YLoc and MultiLoc2 have very different performances between the Yeast Low-Res and High-Res datasets, which could be due to the use of differ-

Table 2.5: Prediction performance(MCC Scores) of individual predictors for the Human dataset

	Cytosol	Mitochondrion	Nucleus	Secretory	Overall Accuracy
YLoc (2010)	0.308	0.546	0.454	0.720	0.628
MultiLoc2 (2009)	0.334	0.451	0.293	0.627	0.581
KnowPred (2009)	0.307	0.048	0.419	0.477	0.514
Subcell (2008)	0.050	0.080	0.122	0.205	0.303
WoLFPSORT (2007)	0.261	0.329	0.277	0.553	0.527
BaCelLo (2006)	0.220	0.439	0.233	0.607	0.54
CELLO (2006)	0.117	0.369	0.234	0.428	0.419
SubLoc (2001)	0.065	0.264	0.162	0.339	0.375
[1]LR	0.362	0.515	0.375	0.712	0.646

[1]LR with all 8 predictors

ent training datasets. For the Human dataset (Table 2.5), the Secretory proteins (which are exclusively Extracellular proteins) are the easiest for YLoc, MultiLoc2, and WoLFPSORT, which may suggest that these proteins have more distinct features such as secretory pathway signals than the Yeast Secretory proteins. As shown in Table 2.1, YLoc, MultiLoc2, and WoLFPSORT all use sorting signals as one of their features. The variation of prediction performance of the individual predictors implies that an ensemble algorithm may be able to integrate their strengths and achieve better overall performance.

2.3.2 Ensemble performance

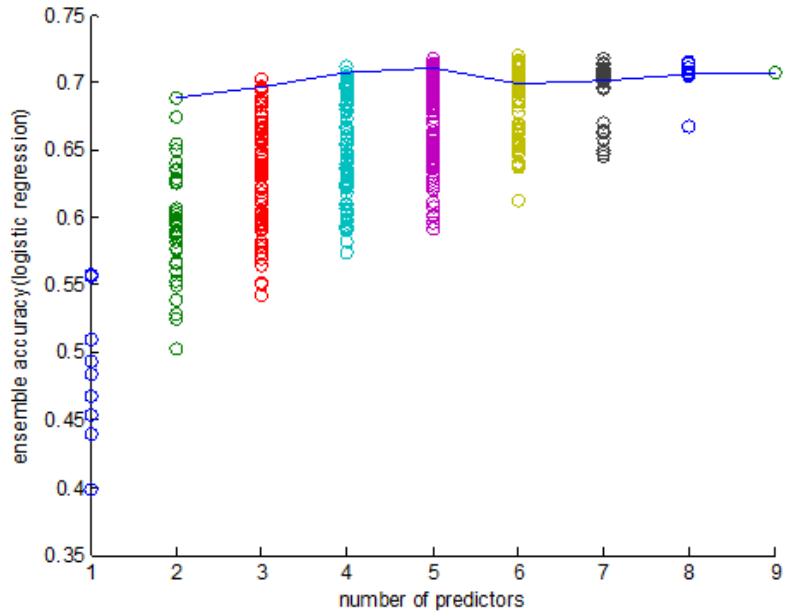
From Table 2.3 to 2.5 we can compare the performances between logistic regression (LR) ensemble algorithms and their element predictors on the three test datasets.

We can see that LR ensemble has better overall accuracy than the best element predictor over the three datasets; for the Yeast Low-Res dataset and Yeast High-Res dataset, LR ensemble have more than 10% improvement over the best element predictors when integrating all available element predictors. However, LR ensemble does not always have the best performance on each compartment. This is because the ensemble training process is to optimize the overall accuracy while performance of certain compartment(s) could be compromised. We can also see that when all of the element predictors failed on certain compartments, such as Golgi and Cell Periphery in the Yeast High-Res dataset, LR ensemble doesn't have any improvement on predicting those compartments.

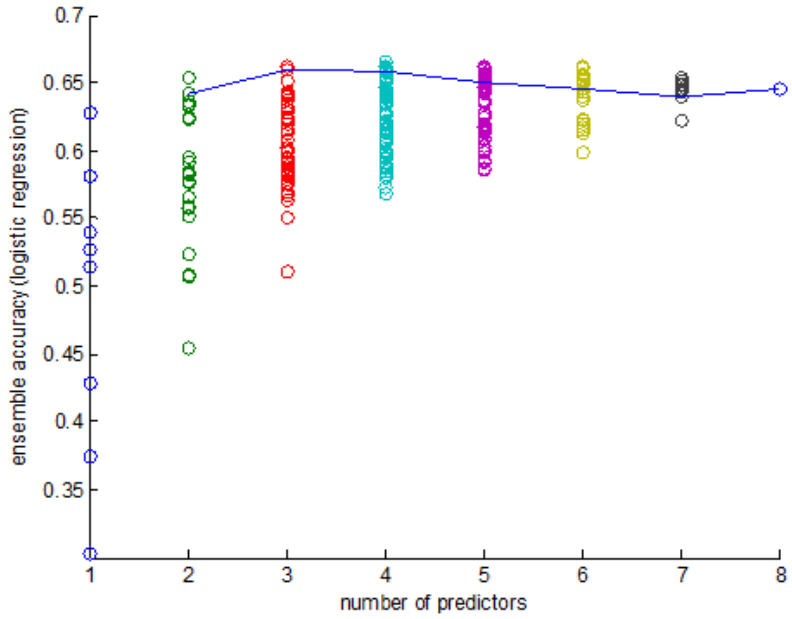
2.3.3 Prediction performance of the optimal ensemble algorithms using exhaustive search

Here we evaluated the prediction accuracy of the logistic regression ensemble algorithm with all combinations of K (K=2...9) predictors using 10-fold cross-validation. Figure 2.1(a) shows the result tested on the Yeast Low-Res dataset. First, we found that by using just three predictors, the ensemble algorithm can achieve comparable performance as using nine predictors. The 3 predictors are NetLoc (PPI), WoLFPSORT and YLoc which cover most of the available features among the predictors. On the other hand, the ensemble algorithm composed of predictors with low coverage of features has poor prediction efficiency. It is also observed that when more predictors were used, the performance discrepancy between the ensemble algorithms based on different predictors became smaller. This indicates that the prediction performance is more reliable as the number of predictors increases.

We also evaluated the ensemble performance on the Human dataset with all combinations of predictors including YLoc, MultiLoc2, WoLFPSORT, CELLO, SubLoc, Subcell, BaCelLo and KnowPred. However, relatively limited accuracy improvement



(a)



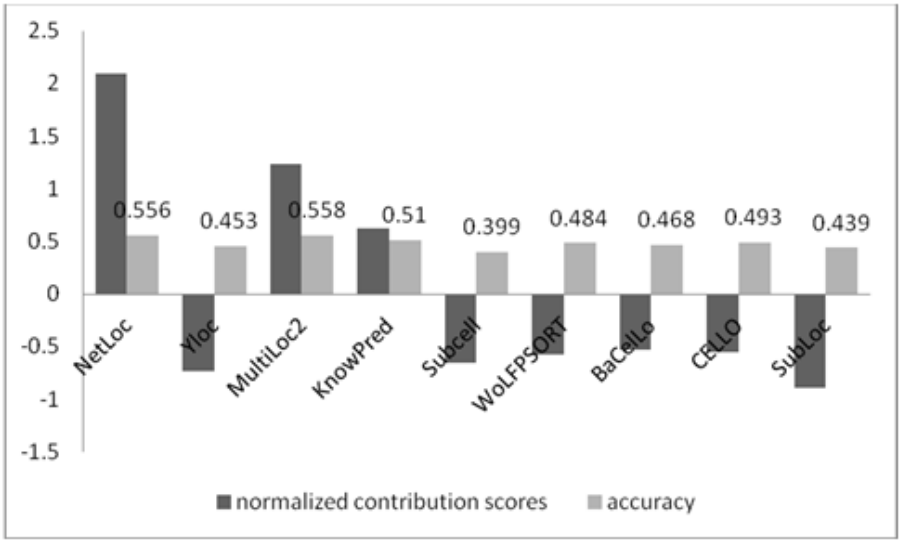
(b)

Figure 2.1: Prediction performance of the logistic regression ensemble methods with K individual predictors selected by exhaustive search. (a) Performance on the Yeast Low-Res dataset, (b) Performance on the Human dataset. Each dot represents one combination of predictors. The number of predictors is annotated on the X axis. The performance of the logistic regression ensemble method is annotated on the Y axis. The dots connected by the line represent the combinations of predictors determined by the minimalist algorithm for different K values.

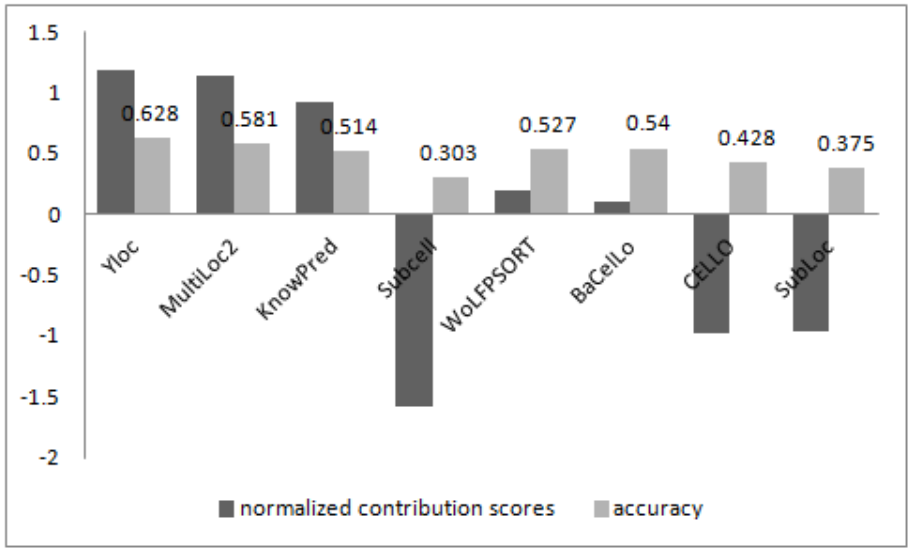
over the best individual predictor has been achieved by the LR ensemble compared to the Yeast dataset. One reason is that the ensemble algorithm for the Yeast dataset includes NetLoc which uses protein-protein correlation network information for localization prediction. This distinctive feature makes it complementary to the other algorithms, which leads to significant performance boosting. Another reason may be that the strengths and bias of different predictors are enlarged or reduced to different degrees on different datasets, which may result in the change of complementary relationship among predictors. The varying complementary relationship thus leads to different prediction accuracy of the ensemble composed of the same set of predictors on different datasets.

2.3.4 Contributions of individual predictors to the ensemble algorithm

To explore the contributions of individual predictors to the ensemble algorithm and their redundant or complementary relationships, we calculated their contribution scores in the ensemble algorithm for the Yeast Low-Res and Human datasets. Nine predictors are available for the Yeast Low-Res dataset and 8 predictors for the Human dataset. Figure 2.2(a) and 2.2(b) show the normalized contribution scores and prediction accuracies of the 9 (8) predictors on the Yeast Low-Res dataset and Human dataset respectively. For the Yeast Low-Res dataset, YLoc2, Subcell, WolfPSORT, BaCelLo, CELLO, and SubLoc all have relatively low contribution scores, which suggests that their predictions are highly redundant with the other predictors' predictions. We also found that the predictors simply using the most common features(amino acids composition) such as CELLO, SubLoc, Subcell, all have relatively low contribution scores, which suggests that the proteins whose localizations can be correctly predicted by these predictors can also be predicted correctly by other predictors. On the other hand, it can be observed that predictors using distinct features such as NetLoc and KnownP have relatively high contribution scores. NetLoc



(a)



(b)

Figure 2.2: Contribution scores of individual predictors. (a) 9 predictors for the Yeast Low-Res dataset, (b) 8 predictors for the Human dataset.

(PPI) has the highest contribution score because it used very different PPI information compared to other predictors, which allows it to correctly predict proteins that other individual predictors cannot. KnowPred applies a sophisticated local similarity method to detect remote sequence homology and therefore might correctly predict some proteins that most of others cannot. Another reason why NetLoc and KnowPred have relatively high contribution scores is that they don't use other common features so they are less likely to make the same wrong predictions like other predictors. For the Human dataset, YLoc, MultiLoc2 and KnowPred have the highest contribution scores while CELLO, SubLoc, and Subcell still have the lowest contribution scores, which suggests that the latter three predictors' correct predictions can be covered by the other component predictors or that they tend to mislead the ensemble algorithm by making majority incorrect predictions. This contribution score analysis can thus be applied to evalute future new protein localization predictors in terms of their unique prediction capability.

2.3.5 Prediction performance of the minimalist ensemble algorithm

To test the performance of our minimalist LR ensemble algorithm with K component predictors, we run the minimalist algorithm to generate the combination of predictors for each K to build the minimalist ensemble algorithms and then tested them on the Yeast Low-Res and Human datasets. The results in Figure 2.1 show that for the LR ensemble method, our minimalist ensemble algorithm can achieve near-optimal performance for any given K value. We also found that using 3-4 individual predictors can obtain near-best performance for all possible K values on the Yeast Low-Res dataset. This means that our minimalist ensemble algorithm can use 1/2 to 1/3 of individual predictors used by existing ensemble algorithms to achieve similar performance while remarkably reducing the computational effort.

To examine the complementary relationships of the selected algorithms in the

ensemble algorithms, Table 2.6 shows the most frequent predictors selected by the minimalist ensemble algorithms during the 10-fold cross-validation and the best combination for each K according to the exhaustive search of the LR ensemble on the Yeast Low-Res dataset. It is interesting to find that NetLoc and WoLFPSORT are the key component algorithms that are selected by the best combination and the minimalist ensemble with different K components. YLoc is the second tier of algorithms selected by the best combination, while MultiLoc2 is the second tier of algorithm selected by the minimalist algorithm. The consistent difference of the selected component predictors between the best combination and the minimalist after the key component algorithms is due to that our minimalist algorithm used greedy and stepwise method to search the optimal K component predictors.

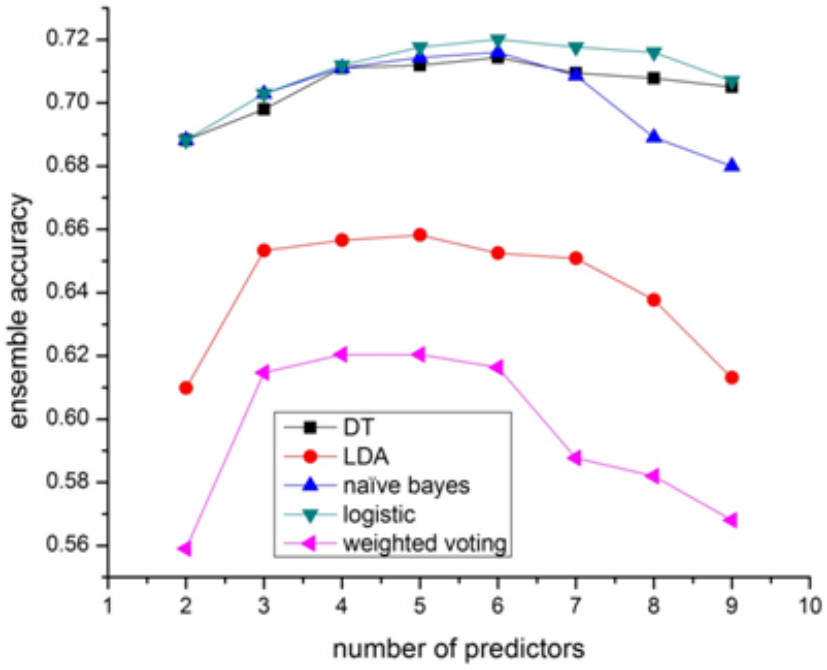
Table 2.6: The most frequent predictors selected by the minimalist algorithm with size of each K (note by M) during the 10-fold cross-validation and the best combination of K predictors (noted by B) according to the exhaustive search result of the logistic regression ensemble on the Yeast dataset

Number of predictors	2	3	4	5	6	7	8
YLoc (2010)		B	B	B	BM	BM	BM
NetLoc (2009)	BM	BM	BM	BM	BM	BM	BM
MultiLoc2 (2009)		M	BM	M	M	M	BM
KnowPred (2008)			M	BM	BM	M	BM
Subcell (2007)						B	B
WoLFPSORT (2006)	BM	BM	BM	BM	BM	BM	BM
BaCelLo (2006)						BM	M
CELLO (2006)				M	BM	BM	BM
SubLoc (2001)				B	B	B	BM

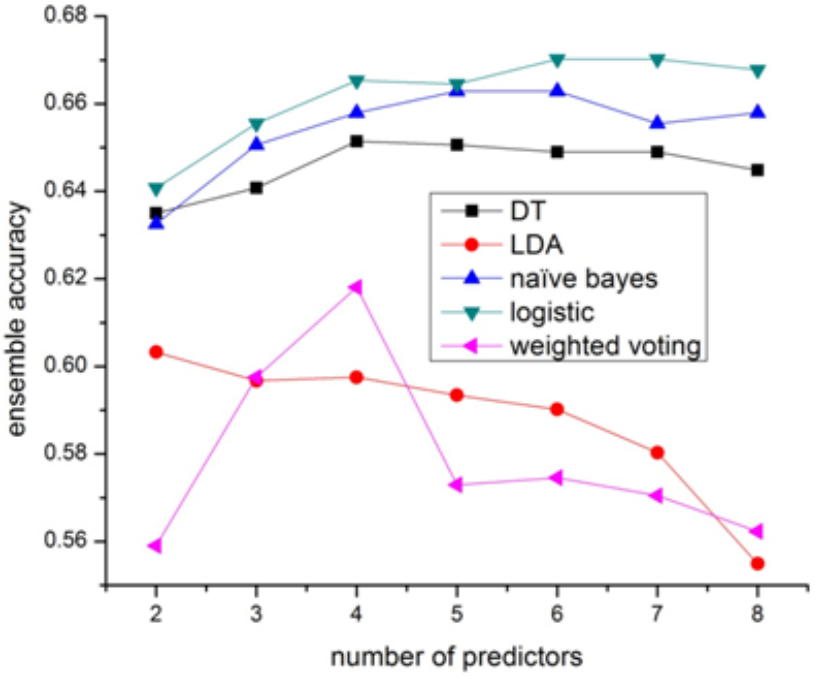
2.3.6 Comparison of computational complexity

The computational complexity of the ensemble involves the effort to collect prediction results from individual predictors either from local software running or from web servers and the total running time. Since most of the predictors are available only via web servers which are sometimes offline, it is desirable to have fewer component predictors. As demonstrated in Figure 2.1, the minimalist algorithm can efficiently find the key component predictors. Since only 4 predictors are needed for the ensemble algorithm to achieve comparable performance of using 9 predictors, about 1/2 to 2/3 amount of computation time to collect prediction results can be saved.

Several ensemble schemes have been proposed for building ensemble localization prediction algorithms, including weighted voting [4] (weight is assigned based on predictor accuracy), LDA [6], and classifiers-based ensemble algorithms such as decision tree (DT) [7]. It is interesting to compare their performance on the genome-wide Yeast and Human datasets. Here we compared their best performance given K individual predictors selected by exhaustive search. As shown in Figure 2.3, weighted voting has the worst performance and its performance degrades dramatically when more individual predictors are included. This is because its prediction can be easily biased by redundant low-performance predictors. LDA ensemble is better than weighted voting because it can assign LDA optimal weights to predictors and avoid the prediction results being biased by low-performance predictors. However, it is still a voting based algorithm which might not be able to capture the rules relating the predictions of predictors to the real locations. For other classifiers-based (such as naive Bayes, decision tree and logistic regression) ensemble methods, they yield better prediction accuracy because these machine learning algorithms can better find and learn the rules between the features (predictions of individual predictors) using supervised learning. For these machine learning ensemble methods, the capability to handle redundancy is essentially the capability to handle over-fitting. As Figure 2.3

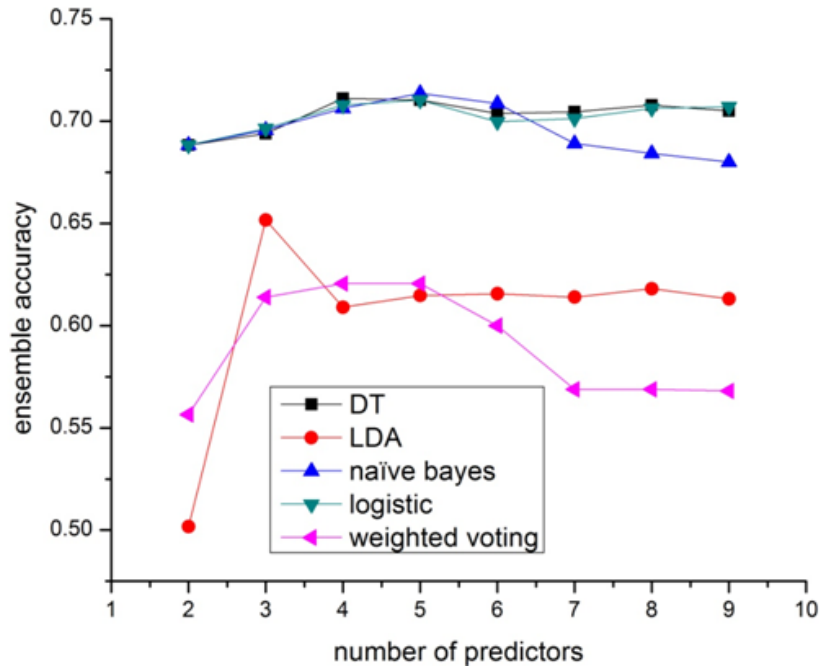


(a)

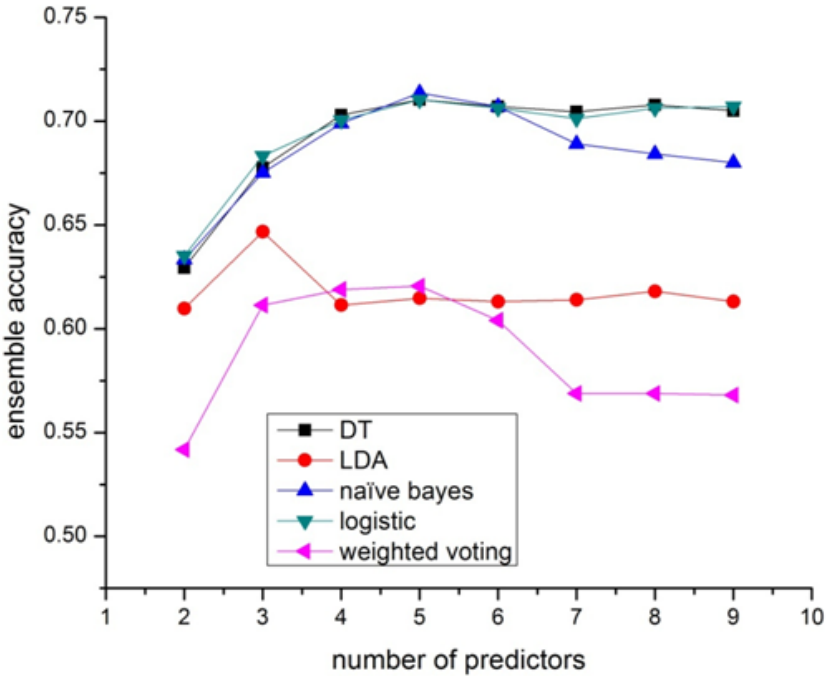


(b)

Figure 2.3: Performance of the best ensemble on the Yeast dataset using different ensemble schemes with K (K = 2..9) predictors selected by exhaustive search. (a) 9 predictors including NetLoc (PPI) (b) 8 predictors without NetLoc (PPI).



(a)



(b)

Figure 2.4: Performance of different ensemble schemes on the Yeast Low-Res dataset with K (k = 2..9) predictors selected by Minimalist algorithm and Top-K accurate method. (a) Different ensemble methods with K (k = 2..9) predictors selected by Minimalist algorithm. (b) Different ensemble methods with K (k = 2..9) predictors selected by Top-K accurate algorithm.

shows, if too many predictors are included, voting based ensemble algorithms such as weighted voting and LDA show the trend of downgrading the performance.

Figure 2.3(a) and 2.3(b) showed the performance of the ensemble algorithms with or without including the PPI based predictor NetLoc. It is observed that ensemble algorithms without NetLoc have much less improvement over the best individual predictors, which means that these ensemble algorithms except weighted voting can automatically take advantage of the unique/beneficial component predictors (such as NetLoc which uses a unique protein-protein interaction features) to improve the performance. From Figure 2.3(b) we also noticed that LDA ensemble's performance could degrade dramatically when too many redundant predictors are included without including predictor(s) with distinct property such as NetLoc.

We also compared the performances of the minimalist ensemble algorithms on the Yeast Low-Res dataset. The result is shown in Figure 2.4(a), which demonstrates similar relationship of the performance for the evaluated ensemble algorithms in Figure 2.3(a) and 2.4(b) shows the performance of the ensemble methods by selecting the top K accurate predictors. We can see that the main peformance difference between the minimalist ensemble and top-K ensemble is when K is less than 4, which means the top 4 accurate predictors can form a very complimentary group. However, top K method is not reliable especially when the predictor with distinct features has relatively low accuracy, or when many included predictors are highly redundant.

2.3.7 Comparison with other ensemble algorithms

There are several published and publicly available ensemble algorithms such as ConLoc [6] and PROlocalizer [5]. ConLoc intergrated 13 different predictors and used LDA as the ensemble scheme. PROlocalizer intergrated 11 different programs to predict localization of animal proteins. We tested ConLoc on our Yeast Low-Res and Human datasets. The results are shown in Table 2.7 and 2.8. It should be noted that

although our datasets are not overlapped with ConLoc ensemble training dataset, the performance result of ConLoc can still be overestimated since we didn't exclude proteins of our datasets that are overlapped with the training datasets of ConLoc's 13 element predictors. To test our minimalist ensemble algorithm, we first collected predictions of ConLoc's 13 element predictors on the Yeast Low-Res and Human datasets and then tested LR ensemble with 10-fold cross-validation. The results (Table 2.7 and 2.8) showed that LR ensemble achieved higher accuracy than LDA based ConLoc on both datasets, which is consistent with our previous experiment results (Figure 2.3(a) and 2.3(b)) alghough ConLoc LDA used a different ensemble training dataset.

To investigate the redundancy among ConLoc's 13 predictors, we applied our minimalist algorithm to select K out of the 13 predictors and tested them on the Yeast Low-Res dataset and the Human dataset. The results (Table 2.7 and 2.8, column 5) showed that for the Yeast Low-Res dataset, using only 4 predictors can achieve equally good performance as using all the 13 predictors. The most frequent 4 predictors selected by our minimalist algorithm during the 10-fold cross-validation are CELLO, Proteome Analyst, PTS1Prowler, and SherLoc. For the Human dataset, using only 3 predictors can achieve better performance than using all the 13 predictors. The most frequent 3 predictors selected by our minimalist algorithm during the 10-fold cross-validation are Proteome Analyst, PTS1Prowler, and SherLoc.

We also tested PROlocalizer which is an integration algorithm based mainly on binary classifiers. However, the server was able to generate prediction results for only 399 out of 1305 proteins in our Human dataset. The overall prediction accuracy of PROlocalizer on those 399 proteins is 0.81 while the standalone predictor YLoc alone has an overall accuracy 0.84 on the same dataset. We argue that it is difficult to construct a reliable protocol-based ensemble algorithm such as PROlocalizer when the predictions of individual predictors are still not reliable leading to accumulation

Table 2.7: Comparison of the performance of ConLoc and Minimalist LR ensemble algorithm with 13 predictors on the Yeast Low-Res dataset

	The best element predictor of ConLoc: SherLoc	ConLoc	[1]LR	[2]LR
Cytosol	0.301	0.441	0.489	0.472
Mitochondrion	0.574	0.622	0.708	0.731
Nucleus	0.341	0.461	0.537	0.541
Secretory	0.533	0.537	0.608	0.605
Overall Accuracy	0.529	0.616	0.696	0.693

[1]LR ensemble with 13 predictors used in ConLoc
[2]LR+minimalist algorithm to select K out of 13 predictors in ConLoc, K=4

Table 2.8: Comparison of the performance of ConLoc and Minimalist LR ensemble algorithm with 13 predictors on the Human dataset

	The best element predictor of ConLoc: Proteome Analyst	ConLoc	[1]LR	[2]LR
Cytosol	0.390	0.414	0.429	0.460
Mitochondrion	0.613	0.628	0.641	0.645
Nucleus	0.463	0.415	0.371	0.392
Secretory	0.754	0.721	0.749	0.758
Overall Accuracy	0.644	0.664	0.689	0.703

[1]LR ensemble with 13 predictors used in ConLoc
[2]LR+minimalist algorithm to select K out of 13 predictors in ConLoc, K=3

of errors along its sequential inference steps. Instead, the machine learning based ensemble methods can learn complementary rules among the predictors to function as a "protocol" to determine protein localization.

2.4 Conclusions

Although many protein localization prediction algorithms have been developed, the prediction performance remains low and the features used to predict localizations are still limited. Ensemble algorithms have shown some promise to take advantage of a

variety of features by combining individual predictors. However, combining as many as possible individual predictors, which is the most common strategy, has the drawback of high running complexity and low availability as well as risk of performance degradation. The result of our minimalist ensemble algorithm showed that it is possible to significantly reduce the number of individual predictors in a given ensemble algorithm while maintaining comparable performance. It is also observed that the best component algorithm set tends to keep predictors with unique features, which indicates that new features are the key to further improve the prediction accuracy for localization prediction. The success of our minimalist ensemble algorithm based on feature selection and logistic regression showed that supervised ensemble algorithms based on machine learning can effectively capture the complex relationships among individual predictors and achieve better performance than the voting methods.

Chapter 3

SeqNLS: Nuclear localization signal prediction based on frequent pattern mining and linear motif attributes

3.1 Background

A nuclear localization signal is a protein peptide bound to carrier proteins for trafficking nuclear proteins into the nucleus. As the most direct evidence for nuclear localization, identification of NLSs can help to elucidate protein functions. However, experimental identification of such signals is costly and currently only a limited number of NLSs have been identified. It is thus desirable to develop algorithms for computational prediction of NLSs. Several NLS prediction methods have been developed such as PSORT II [19], PredictNLS [20], NLStradamus [21], cNLS Mapper [22], and NucImport [23]. PSORT II predicts NLSs based on sequence patterns implemented as three simple rules according to the classification of NLSs [24]; the rules are mainly clusters of basic amino acids K and R and gaps between the clusters. PredictNLS predicts NLSs based on 194 potential NLS motifs, which are derived from 114 experimentally verified NLSs with a silico mutagenesis approach. Nguyen Ba et. al. [21] found that NLSs tend to have similar residue frequency distributions which are different from that of background residues. Their NLStradamus algorithm detects NLSs by using a simple two-state or four-state HMMs to accommodate the frequency variations. cNLS Mapper estimates classical NLS (cNLS) functionality of a peptide by calculating sum of the functional contribution of each residue in the peptide ac-

cording to the activity-based profiles, which are obtained from the systematic amino acid-replacement analyses in budding yeast. NucImport builds a Bayesian network to predict nuclear localization by incorporating various attributes related to the nuclear importing. If a protein is predicted as a nuclear protein, the location of its NLS is predicted as the segment in the protein with the highest cNLS score in the inferred cNLS class based on the dependencies with other attributes in the Bayesian network.

These five NLS prediction methods have achieved different degrees of success. However, their prediction performance is still far from being sufficient to assist biologists to discover putative NLSs in protein sequences of interest. Each of them has their weakness. Although a great portion of NLSs can be covered by the rules used in PSORT II to detect NLS, quite a few patterns covered by the rules exist in peptides which do not contain NLSs, leading to a high false positive rate or low prediction precision. The sensitivity of the PredictNLS algorithm depends on the number of NLS motifs it used, which has been extended by introducing the potential NLS motifs generated using in-silico mutagenesis analysis. But they are still too specific and couldn't effectively accommodate NLS variability [21]. The performance of the NLStradamus algorithm depends on its assumption that NLSs have certain residue distributions. However, many NLS instances in our testing datasets have shown very different residue frequencies. Both cNLS mapper and NucImport algorithms are developed based on the characteristics of cNLS. However, approximately 43% of proteins localized to the nucleus may use other transport mechanisms other than the classical nuclear import pathway according to Lange et al [25].

One of the challenges of NLS prediction is that functional NLSs are not defined [70]. Many NLSs are short peptides that occur regularly in non-nuclear proteins. In fact, NLS is one type of linear motifs as defined in the database of eukaryotic linear motifs [71]. Linear motifs are short stretches of residues which are highly involved in cell signaling and regulating. To adapt to the fast fine-tuning cell regulatory process,

certain characteristics of linear motifs have thus evolved and might have contributed to NLS variability: only a few residues within a linear motif are functionally important, and mutation of a single residue can switch on/off the functionality [72, 73]. The nature of shortness, flexibility and sensibility provides linear motifs evolutionary plasticity to form a functional unit and fine-tune cell singling network over short evolutionary distances, which, however, increases the difficulties in computational identification of linear motifs such as NLSs.

In the past decade, many computational approaches have been proposed to discover linear motifs. There are two categories of the methods [72]: one is supervised methods aiming to identify new instances of known linear motifs in protein sequences [71, 74–80]; the other is de novo methods for discovering new linear motifs [81–84]. The challenge of the former is to discriminate between true and false positive matches. Most of such prediction algorithms take advantage of the special attributes of linear motifs [85] to remove false positive matches that are unlikely to be functional linear motifs. The latter de novo linear motif discovery algorithms [81, 82, 84] are usually based on the enrichment analysis of candidate motifs integrated with disorder prediction and evolutionary conservation. Since NLS is one type of linear motifs, the framework of the first category may apply to predicting NLSs. However, despite the availability of a number of NLS motifs [86, 87], they are either too specific [21] or they only target a specific pathway of NLSs. To cover more NLSs, we need a new approach to utilize linear motif attributes.

In this chapter, we propose a novel algorithm for NLS prediction based on sequential-pattern mining and linear motif scoring. Our strategy is first to detect potential NLS candidates using the sequential-pattern mining method, which are then scored in terms of their likelihood of being (part of) NLS based on their sequence and linear motif features. The qualified candidate motifs will then be combined into NLS predictions.

3.2 Materials and Methods

3.2.1 Training and Testing dataset

We used 114 experimentally determined NLSs from NLSdb [86] as the source of the positive training dataset for sequential pattern mining. Two NLSs without a specific form in amino acid sequence and a reference citation were removed. 94 out of 112 were real NLSs of which the parent proteins could be found, while the rest 18 were either synthetic NLSs or NLSs of which the parent proteins couldn't be found. We then removed the redundant NLSs in order to avoid non-functional residues being enriched in the positive training dataset: given a NLS A, the redundant NLSs to A are defined as NLSs whose parent proteins are highly homologous to the parent protein of NLS A and are overlapped with NLS A in the alignment of their parent proteins. To remove redundant NLSs, Blastclust with 90% sequence identity and 90% sequence coverage was applied on the parent proteins of the 94 NLSs. If multiple NLSs were overlapped in the alignment of their parent proteins which were in the same cluster, then only one of the NLSs was kept; 4 out of the 94 NLSs were thus removed. In the end, 108 experimentally verified NLSs were left in our positive training dataset for sequential-pattern mining. We then collected 2238 non-nuclear proteins from the BaCello dataset [59], from which 26772 non-overlapped peptides of length 40 were randomly sampled for the negative training dataset for sequential-pattern mining. The length 40 was determined because it approximated the longest NLSs in the positive training dataset. To prepare the training dataset for linear motif scoring (to be defined below), the 90 NLSs with known parent proteins used in the training dataset of sequential-pattern mining were used as the positive training dataset. For each of the 90 NLSs, a random amino acid segment of the same length in the same parent protein which was not overlapped with any annotated NLS was collected to produce the negative dataset.

We prepared two independent testing datasets according to the species of the NLS source proteins for evaluating the NLS predictors: 1) The Yeast NLS dataset; 2) The Hybrid NLS dataset of which the parent proteins are from different species. The Yeast dataset was prepared based on the dataset used in NLStradamus [21]. The Hybrid dataset was collected by searching annotated NLSs from literature published after 2010. All NLSs in the testing datasets redundant to NLSs in the training dataset (90 NLSs with known parent proteins) were removed, and redundant NLSs in the testing dataset itself were also removed. In the end, the Yeast dataset contains 50 NLSs from 41 proteins, and the Hybrid dataset contains 73 NLSs from 53 proteins. Both datasets are provided in the supplementary file (Table S1 and Table S2).

3.2.2 Overview of the proposed algorithm

Our SeqNLS algorithm is developed based on the following observations of NLSs: 1) most known NLSs are composed of a sequence of well-conserved segments of amino acids with variable-length gaps. This is because a set of NLSs binding to the same binding pockets usually share such patterns due to the geometrical or physical-interaction constraints at the binding interface. Such sequential patterns are thus over-represented among these NLSs; 2) similar to other linear motifs, NLSs usually occur in the disordered regions of the protein sequences; 3) NLSs for different pathways may be different. Our algorithm for NLS prediction can be divided into two steps: 1) mining NLS sequence patterns from experimentally verified NLS instances and then predicting NLS candidates on query sequence(s); 2) scoring candidate NLSs based on sequence and linear motif scoring and applying local conservation masking. Our sequential-pattern mining method is motivated by the fact that diversity among the experimentally verified NLSs has hampered the discovery of NLS motifs due to a limited number of NLS instances [87, 88]. SeqNLS addresses this issue by using a more general motif model: the sequential patterns.

3.2.3 Sequential-pattern based prediction of NLSs

In our method, sequential pattern mining is used to extensively collect potential NLS segments/building blocks, which are then used to detect potential NLS segments in query sequences.

NLS sequential-pattern mining

Figure 3.1(a) shows the flow chart of NLS sequential-pattern mining on a training dataset. We first define a segment of amino acids as a word, and a set of words in sequential order as a word-list; the NLS sequential patterns are thus defined as word-lists over-expressed in a set of NLSs (positive training dataset) against a set of peptides not overlapped with any NLS (negative training dataset). The number of different word-lists within the positive training dataset is too large while many of them are redundant; therefore, we limit the search space of word-lists as frequent word-sets within the positive training dataset, which can effectively reduce the search space and maintain the diversity of word-lists; the frequent word-set is defined as a word-set with support count no less than 3 within the positive training dataset and set size not larger than 4. For example, if there are 12 NLSs in the positive training dataset containing the word-set AT, KK, the word-set AT, KK is a frequent word-set since its support count is 12 and the set size is 2. We apply the frequent item-set mining algorithm [89] to collect all the frequent word-sets within the positive training dataset in step 1; the word-lists are obtained by permuting each of the frequent word-sets, and the corresponding support counts in the positive training dataset are then collected in step 2; in step 3, all the word-lists are scored according to their corresponding occurrences in the positive and negative training datasets to measure their enrichment. The enrichment score is defined as follows:

$$E_S = \log(N_{P1}/N_P)/((N_{B1}+1)/N_B)$$

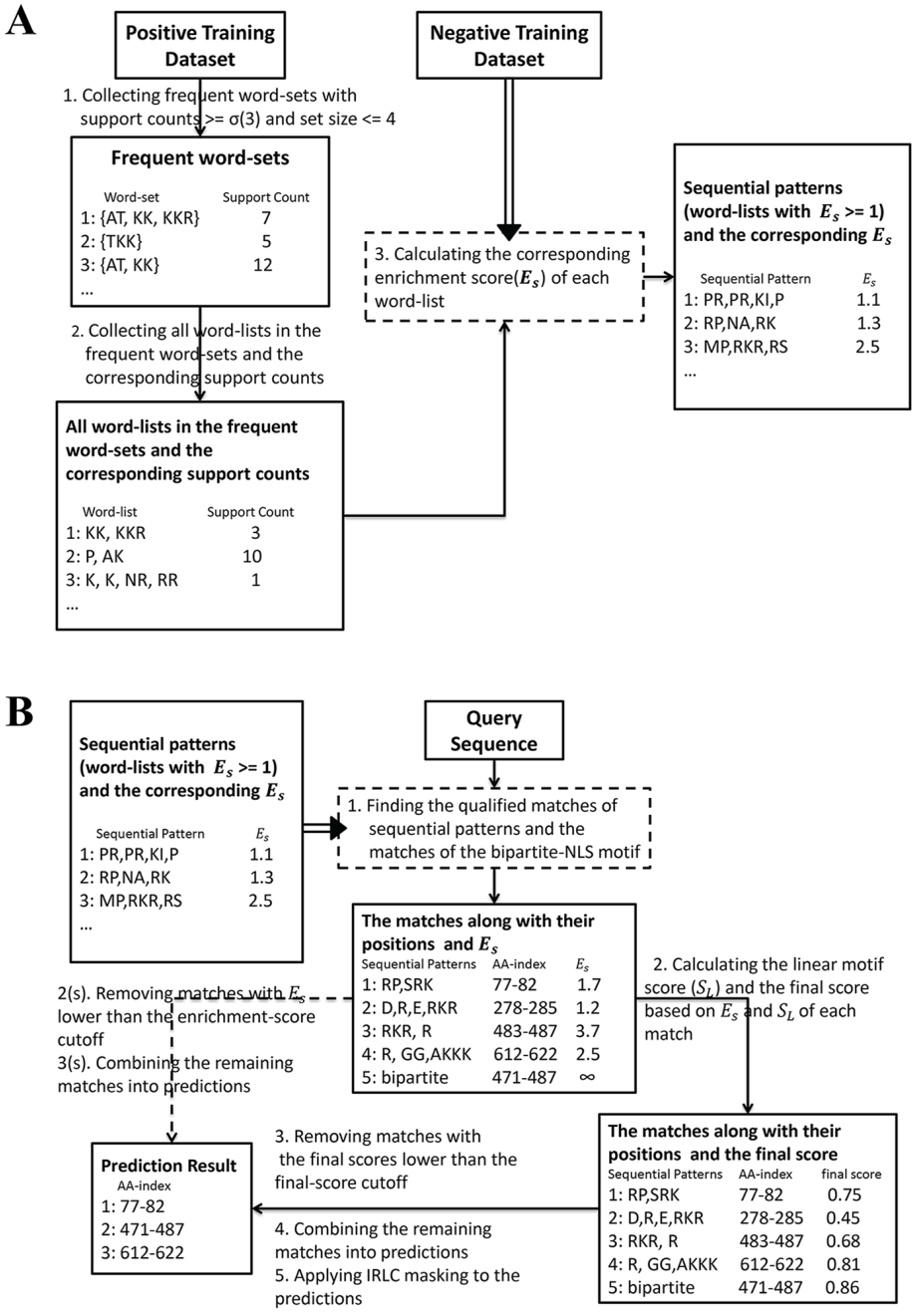


Figure 3.1: The flow charts of predicting NLS. (a) The flow chart of mining the sequential patterns. (b) The flow chart of predicting NLS on a query sequence; the dashed line corresponds to the sequence-based predictor, and the other branch using linear motif scoring refers to the integrated prediction algorithm.

where E_S is the enrichment score, N_P is the number of NLSs in the positive training dataset while N_{P1} is the number of NLSs in the positive training dataset containing the word-list, and N_B is the number of peptides of length 40 that are not overlapped with NLS in the negative training dataset while N_{B1} is the number of peptides in the negative training dataset that contain the word-list; N_P is 108 and N_B is 26772 according to our training dataset. The enrichment score E_S is essentially a measure of over-representation for the word-lists in the training NLSs relative to the non-NLS peptides. The word-lists with E_S not lower than a default threshold 1.0 are collected as the sequential patterns, which will then be used to detect segments which are likely to be (parts of) a NLS in a query sequence.

Detecting potential NLS segments by using the NLS sequential patterns

The process to detect potential NLS segments by using the collected NLS sequential patterns is illustrated in Figure 3.1(b). First, the collected sequential patterns are used to find qualified matches in the query sequence, which are defined as the matches of the sequential patterns in the query sequence with each gap between the words no longer than two amino acids. The reason to limit the length of the gaps is to maintain the statistical significance of the sequential-pattern matches since it is much more likely to have words in a specific order by chance when long gaps are allowed. These qualified matches are recognized as potential NLS segments in our algorithm, of which E_S is a measure of the significance of these potential NLS segments to be true NLS. In Figure 3.1(b), the dashed line corresponds to our sequence-based predictor, and the other branch using linear motif scoring refers to our integrated prediction algorithm.

3.2.4 Incorporation of bipartite-NLS motifs

Our SeqNLS algorithm does not make any assumptions over the type of the predicted NLSs. However, to improve the prediction performance, a bipartite-NLS motif is incorporated in SeqNLS to increase the sensitivity of detecting bipartite NLSs. Bipartite NLSs are a class of classical NLS usually composed of two clusters of basic amino acids separated by a gap of 10-12 residues [90, 91] while longer gaps are also possible [92]. Bipartite NLSs are very common as it was approximated that 25.8% of proteins localized to the nucleus contain putative bipartite NLSs [25]. Several consensus patterns of bipartite NLSs have been defined such as $(K/R)(K/R)X_{10–12}(K/R)_{3/5}$ [93], $KRX_{10\text{-}12}KRRK$ [94], and $KRX_{10–12}K(K/R)(K/R)$ or $KRX_{10–12}K(K/R)X(K/R)$ [87], where $(K/R)_{3/5}$ represents any 5 consecutive amino acids having at least three of either lysine or arginine. Since bipartite NLSs have long gaps between the two words, they may not be detected by our sequential-pattern mining method. Therefore, we included a bipartite-NLS motif $(K/R)(K/R)X_{10}(K/R)_{3/5}$, which is also used to predict bipartite NLS in PSORTII, to complement the motifs mined from the training NLSs. As shown in Figure 3.1(b), when detecting potential NLS segments, our algorithm also collects the matches of the bipartite-NLS motif in addition to the qualified matches of the sequential patterns. The matches of the bipartite-NLS motif were found usually more reliable than the matches of sequential patterns according to our experiment result. Therefore the enrichment score of the matches of the bipartite-NLS motif is set as an arbitrarily large value which will never be lower than the enrichment-score cutoff as defined in the next paragraph.

3.2.5 Predicting NLS based on sequence features only: sequence-based predictor

Given a query sequence, the extracted sequential patterns along with the bipartite-NLS motif are used to scan it for matches. Those matches with E_S score lower than a pre-defined cutoff will be removed (the matches of the bipartite-NLS motif will

never be removed). The remaining matches will then be combined using a merging procedure: every two overlapped matches are merged into one match of which the boundaries are defined as the union of the overlapped matches. The merging process will continue until all the matches are not overlapped. The resulting matches will be the output of the sequence-based NLS predictor.

3.2.6 Linear motif scoring

To further improve the performance of NLS prediction, we developed a linear motif-scoring scheme to remove the false positives of the matches as obtained above based on the linear motif attributes. NLSs are one kind of linear motifs, which are found to predominantly occur in disordered regions [85, 95]. One possible reason is that disordered regions can provide linear motifs unstructured interfaces to adapt to the interacting partner with higher flexibility. In addition, evolutionary plasticity inherent to disordered regions increases the likelihood of evolving linear motifs [85]. To exploit this preference of linear motifs, we used PrDOS [96], one of the best-performing disorder predictors according to CASP9 [97], to predict disorder scores for all residues in the query sequence. Given a predicted amino acid segment, the median disorder score of residues within the segment is defined as the disorder score of the predicted peptide.

Another factor to estimate the likelihood of linear motifs is residue accessibility, which is required for linear motifs to function; deeply buried residues are less likely to interact with the partner proteins [98]. In our experiments, NetSurfP [99] was used as the residue-accessibility predictor, and the relative surface area (RSA) was used as the measure of residue accessibility. Given a predicted amino acid segment, the median RSA score of residues within the segment is defined as its RSA score. Our linear motif-scoring scheme is implemented by estimating the probability of being NLS for a given peptide. We call this probability as the linear motif score (S_L).

It is calculated by building a Support Vector Machine (SVM) classifier based on the aforementioned linear motif attributes, whose output is the probability of an input amino acid segment belonging to the NLS class. We collected 90 NLSs and 90 non-NLS peptides (mentioned in the section "Training and Testing dataset") as the positive and negative training datasets for the SVM. The linear motif attributes including the PrDOS disorder score and the NetSurfP RSA score were used as the features. The SVM classifier was trained using the LIBSVM package [100] with the radial basis function as the kernel, and the probability of being NLS for a given input peptide was obtained by calculating the probability estimation of LIBSVM.

3.2.7 Predicting NLS based on sequence and linear motif scoring: SeqNLS, the integrated predictor

Our SeqNLS algorithm works by sequential-pattern mining and matching plus linear motif scoring. First, it collects the matches of the sequential patterns and the bipartite-NLS motif in the query sequence. Next, all the matches of the sequential patterns and the bipartite-NLS motif will be estimated the probability of being NLS by linear motif scoring. The respective linear motif score will then be combined with the corresponding enrichment score to generate the final score. The matches whose final scores lower than a predefined cutoff will be removed. To combine the enrichment score and the linear motif score, we defined the normalized enrichment score which has the same scale as the linear motif score (between 0 and 1). According to our experiment result, we found that when the enrichment-score cutoff is over a certain threshold E_K, the prediction precision cannot be improved by further increasing the cutoff. The normalized enrichment score is thus defined according to the following formula:

$$Normalized(E_S) = \begin{cases} 1 & \text{if } E_S \geq E_K \\ (E_S - Minscore)/(E_K - Minscore) & \text{Otherwise} \end{cases}$$

where Normalized(E_S) represents the normalized enrichment score, and Minscore represents the minimal possible score of E_S, which is 1 according to our setting since only sequential patterns with E_S greater or equal to 1 are collected. The final score will then be calculated according to the following formula:

$$\text{The final score} = \begin{cases} \alpha \times Normalized(E_S) + (1-\alpha) \times S_L & \text{if match is from the bipartite NLS motif} \\ \alpha \times Normalized(E_S) + \beta \times (1-\alpha) \times S_L & \text{Otherwise} \end{cases}$$

It should be noted that the SVM model of calculating S_L is trained to discriminate between NLSs and peptides not overlapped with NLS; however, those true positive matches, which are matches overlapped with NLS according to our definition, do not always have accurate NLS boundaries; the more accurate the NLS boundaries of the true positive matches are, the more reliable their S_L will be. In the formula, S_L of the sequential-pattern matches is multiplied by a weighting factor β (smaller than 1) because we found that the true positive matches of the bipartite-NLS motif generally have more accurate NLS-boundaries in terms of residue-level accuracy. In our study the optimal α and β are set as 0.8 and 0.6 respectively.

3.2.8 IRLC-masking

Due to the short and degenerate nature of linear motifs, the evolutionary conservation of linear motifs cannot be well represented by simple sequence-alignment models. Davey et al [101] proposed the relatively local conservation (RLC) score, which measures the conservation of residues relative to their neighboring regions. They applied RLC masking to remove residues unlikely to be functional residues within linear motifs, based on the rationale that functional residues should be more conserved than

the neighboring regions. While RLC masking has been used to remove false positive matches of known linear motifs [101], it is not an appropriate method to remove false positive NLS predictions due to the fact that those true positive NLS predictions, unlike the true positive matches of other linear motifs, do not always have accurate NLS boundaries and may cover non-functional residues while wildcard positions are not known. Therefore, we proposed the inverse relative local conservation (IRLC) scheme to remove false positive NLS predictions based on the following rationale: since linear motifs are more conserved than their flanking residues, the chance to have a flanking residue which is much more conserved than the residues within the linear motif should be very small.

To evaluate IRLC, we first define M as the mean conservation score of N residues within a predicted NLS:

$$M = \frac{1}{N} \sum_{i=1}^{N} C_i$$

where C_i is the conservation score representing the degree of conservation of a residue in position i of the predicted NLS; C_i can be calculated by any suitable scoring metric, while in our experiment, position specific scoring matrix (PSSM) was used to evaluate residue conservation; the conservation score of a residue in the position i' of a sequence was obtained from the corresponding column of the residue in the i'-th row of the PSSM of the sequence. The PSSM of each query sequence was generated by three iterations of PSI-BLAST [102] searches against NCBI non-redundant database with the BLOSUM62 substitution matrix and E-value threshold of 0.001. Second, we define $IRLC_j$ as the IRLC score for a flanking residue j:

$$IRLC_j = (C_j - M)/\sigma$$

where the flanking residues are defined as the residues within 5 amino acids away

from the predicted NLS, and σ represents the standard deviation of the conservation scores of all the residues in the sequence. The IRLC score for a NLS prediction can thus be defined as:

$$IRLC = \max_j IRLC_j$$

A NLS prediction will be determined as a false positive prediction if its IRLC score is higher than some threshold value T. The rationale is that if there is any residue in the flanking region that is much more conserved than the average conservation score of the region of interest, it is less likely that the region of interest represents a functional NLS since it contradicts the property of relative local conservation of linear motifs.

3.2.9 Performance evaluation

To evaluate NLS prediction performance, a NLS prediction is considered a hit if the prediction is overlapped with at least one annotated NLS in the testing dataset otherwise it is labeled as a miss. Three performance metrics are defined to evaluate NLS prediction performance as follows:

$$\text{precision} = N_{hits}/(N_{hits} + N_{miss})$$

$$\text{recall} = N_{hits}/N_{nls}$$

$$\text{F1 score} = 2 \times \text{precision} \times \text{recall}/(\text{precision} + \text{recall})$$

where N_{hits} is the number of hits, N_{miss} is the number of misses, and N_{nls} is the number of NLSs in the testing dataset. In addition, we introduced the amino acid level performance coefficient [103] (aPC) to evaluate the amino acid-level accuracy of a predicted peptide overlapped with NLS. The aPC is defined as follows:

$$aPC = aTP/(aTP + aFN + aFP)$$

where aTP represents the number of amino acids of a predicted NLS that are overlapped with the true NLS; aFP represents the number of amino acids of a predicted NLS that are not overlapped with the true NLS; aFN represents the number of amino acids of the true NLS that are not overlapped with the predicted NLS. In our evaluation, the mean aPC of all the true positive predictions (Mean aPC) is defined to evaluate the amino acid level accuracy of a predictor.

3.3 Results and Discussion

Table 3.1: The prediction performance of the sequence-based predictor with different enrichment-score cutoffs with and without incorporating the bipartite-NLS motif on the Yeast dataset

[1]Enrich		1.0	1.2	1.4	1.6	2.0	2.3	Bipartite*
Precision		0.212	0.311	0.458	0.564	0.6	0.6	0.667
	+B	0.204	0.303	0.427	0.547	0.613	0.63	
Recall		0.8	0.66	0.6	0.42	0.12	0.06	0.32
	+B	0.8	0.68	0.62	0.56	0.38	0.34	
F1 score		0.335	0.423	0.519	0.482	0.2	0.109	0.432
	+B	0.325	0.413	0.505	0.554	0.469	0.442	
Mean aPC		0.453	0.413	0.412	0.443	0.49	0.442	0.805
	+B	0.554	0.563	0.607	0.645	0.736	0.788	

[1]Enrichment-score cutoff

*Predictions with only the bipartite-NLS motif: $(K/R)(K/R)X_{10}(K/R)_{3/5}$

3.3.1 Performance of the sequence-based NLS predictor

We applied the sequence-based predictor to the Yeast and Hybrid datasets, and the result is shown in Figure 3.2. It shows that when the enrichment-score cutoff is set higher, the precision of the predictor increases. This is because the matches of the

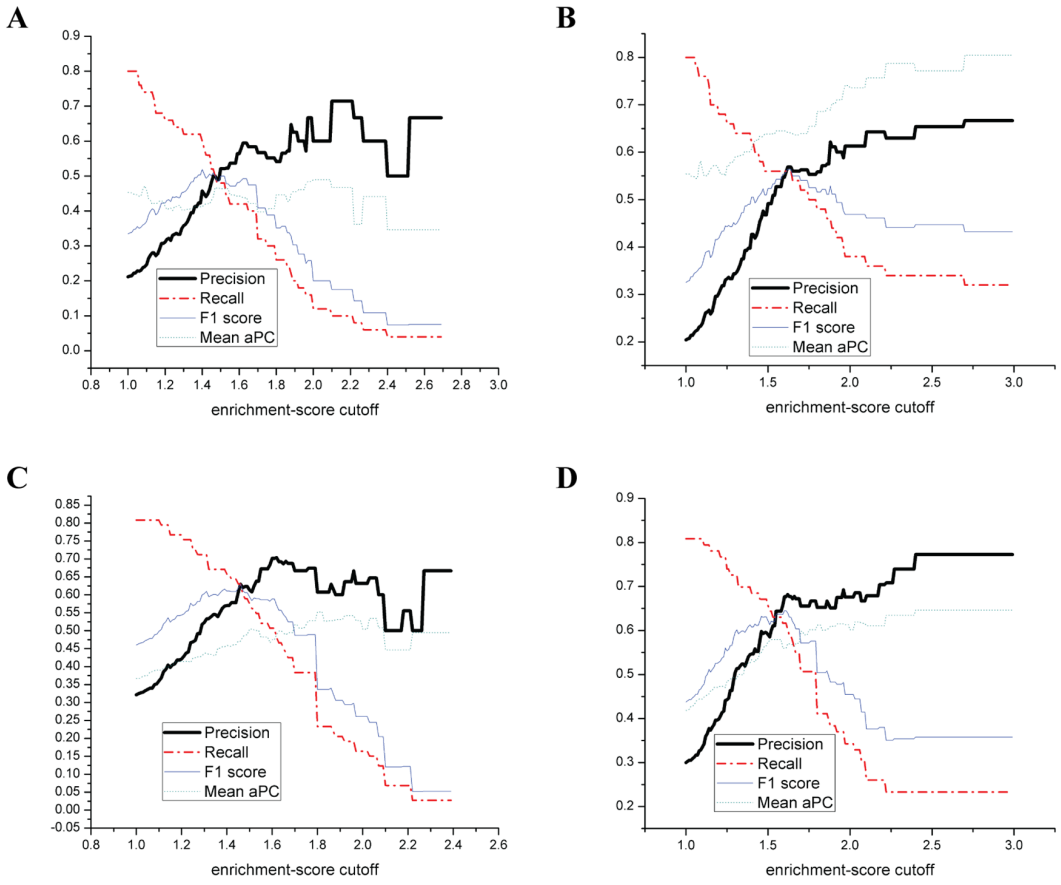


Figure 3.2: The prediction performance of the sequence-based predictor. (a) The Yeast dataset; the bipartite-NLS motif is not incorporated. (b) The Yeast dataset; the bipartite-NLS motif is incorporated. (c) The Hybrid dataset; the bipartite-NLS motif is not incorporated. (d) The Hybrid dataset; the bipartite-NLS motif is incorporated.

Table 3.2: The prediction performance of the sequence-based predictor with different enrichment-score cutoffs with and without incorporating the bipartite-NLS motif on the Hybrid dataset

[1]Enrich		1.0	1.2	1.4	1.6	2.0	2.3	BiPartite*
Precision		0.322	0.421	0.57	0.702	0.632	0.667	0.77
	+B	0.3	0.399	0.546	0.677	0.676	0.739	
Recall		0.808	0.767	0.658	0.507	0.164	0.027	0.23
	+B	0.808	0.781	0.685	0.616	0.342	0.233	
F1 score		0.46	0.544	0.611	0.589	0.261	0.053	0.358
	+B	0.438	0.528	0.608	0.645	0.455	0.354	
Mean aPC		0.367	0.416	0.46	0.475	0.504	0.494	0.646
	+B	0.418	0.473	0.534	0.56	0.601	0.634	

[1]Enrichment-score cutoff

*Predictions with only the bipartite-NLS motif: $(K/R)(K/R)X_{10}(K/R)_{3/5}$

sequential patterns with the higher enrichment score are more significant and thus are more likely to be part of NLS. However, in Figure 3.2(a) and 3.2(c), it can be shown that for both the Yeast dataset and the Hybrid dataset, when the enrichment-score cutoff is higher than 1.62, no obvious precision improvement can be obtained by further raising the cutoff. We thus set E_K as 1.62 in our experiment. In the meantime, recall decreases with the increase of the enrichment-score cutoff. This is because fewer matches can meet the higher enrichment-score cutoff, and thus fewer annotated NLSs can be covered by the matches. The performance of the predictor incorporated with the bipartite-NLS motif is shown in Figure 3.2(b) and 3.2(d). It was found that precision can be further improved by setting a higher enrichment-score cutoff even when the cutoff is higher than 1.62 (E_K). It implies that the bipartite-NLS motif is a more reliable NLS pattern than the mined sequential patterns; by setting the higher enrichment-score cutoff, the proportion of the sequential-pattern matches will become smaller, and the matches of the bipartite-NLS motif will dominate prediction performance when the enrichment-score cutoff is much higher than E_K. To evaluate the performance of the bipartite-NLS motif in NLS prediction, we evaluated the performance of the sequence-based predictor using only the bipartite-NLS motif in

Table 3.1 and Table 3.2 (last column). It was shown that the predictor using only the bipartite-NLS motif has high precision on the both datasets: 0.667 on the Yeast dataset and 0.77 on the Hybrid dataset. It also has very high residue-level accuracy: the Mean aPC is 0.805 and 0.645 on the Yeast dataset and the Hybrid dataset respectively while the Mean aPC of most other NLS predictors is around 0.4 to 0.5. The high precision of the bipartite-NLS motif based predictor is probably due to the high specificity of the bipartite-NLS motif pattern. However, the recall of this method is only 0.32 and 0.233 respectively on the Yeast dataset and the Hybrid dataset.

To evaluate if the bipartite-NLS motif can help to improve the sequence-based predictor, the prediction performance of the sequence-based predictor with or without incorporating the bipartite-NLS motif is shown in Table 3.1 and Table 3.2. It is shown that recall can be improved on both the Yeast and Hybrid datasets after incorporating the bipartite-NLS motif. Improvement on recall depends on the enrichment-score cutoff: when the enrichment-score cutoff is lower, more bipartite NLSs in the testing datasets could be partially covered (overlapped) by the sequential-pattern matches, and thus improvement on recall is smaller. Alternatively, when the cutoff score is higher than 1.6, the incorporation of the bipartite-NLS motif significantly improves recall. Besides, the Mean aPC can be significantly improved by incorporating the bipartite-NLS motif: when the enrichment-score cutoff is set as 1.6, the Mean aPC can be improved from 0.443 to 0.645 on the Yeast dataset and from 0.475 to 0.56 on the Hybrid dataset. Improvement on the Mean aPC also depends on the enrichment-score cutoff: when the enrichment-score cutoff is lower, more bipartite NLSs in the testing dataset are likely to be overlapped with the matches and thus improvement on the Mean aPC by incorporating the bipartite-NLS motif is less obvious. In addition, improvement on both recall and the Mean aPC by incorporating the bipartite-NLS motif also depends on the ratio of bipartite NLSs in the testing datasets, which explains why the improvement on the Yeast dataset is greater than that of the Hybrid

dataset. From Table 3.1 and Table 3.2, we also found that when the enrichment-score cutoff is set as 1.0, 80% of the NLSs can be covered by our sequential-pattern matches for both the Yeast and Hybrid datasets. This indicates that our sequential patterns with the enrichment score higher than 1.0 cover 80% of NLSs, which can be used in searching potential NLSs extensively.

3.3.2 Linear motif attributes of NLS

Here we evaluate the discriminative capacity of linear motif attributes for NLS identification. Figure 3.3(a) shows the disorder propensity of NLSs: the mean PrDOS disorder score of the 90 training NLSs is 0.632 while the mean PrDOS disorder score of the 90 peptides not overlapped with NLS is 0.386. The disorder propensity of NLSs is clearly shown by the peak at index 0, while no such preference exists for the peptides not overlapped with NLS. Figure 3.3(b) shows the RSA propensity of NLSs: the mean NetSurfP RSA score of the 90 training NLSs is 0.393, while the mean NetSurfP RSA score of the 90 peptides not overlapped with NLS is 0.299. The preference of NLSs for higher RSA is also observed by the peak at index 0, while no such preference exists for the peptides not overlapped with NLS. Compared to the disorder propensity, the RSA propensity of NLSs is less significant since the difference of the mean attribute value between NLSs and peptides not overlapped with NLS is 0.094 for NetSurfP RSA, while it is 0.246 for PrDOS disorder (the PrDOS disorder score and the NetSurfP RSA score both have the same scale 0-1).

To further investigate the discriminative capacity of these attributes, we first used each of the attributes to build a single-feature binary classifier in which the prediction is based on the cutoff of the attribute value. The ROC curves of the binary classifiers are plotted in Figure 3.4. As shown in the figure, the AUC values for the PrDOS disorder score and the NetSurfP RSA score are 0.783 and 0.69 respectively. This suggests that PrDOS disorder and NetSurf RSA are both useful features to discriminate

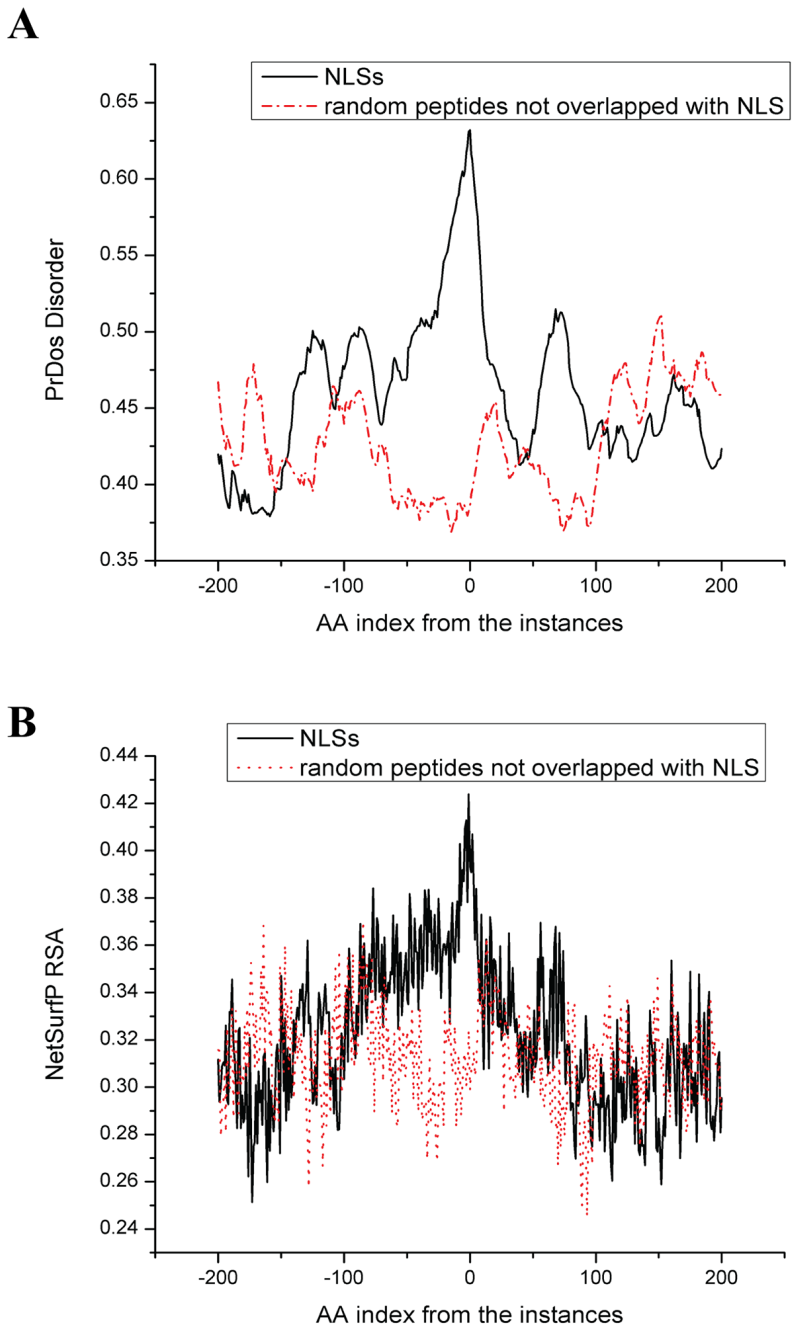


Figure 3.3: The linear motif attributes of NLSs. (a) PrDOS disorder scores of the 200 residues either side of the annotated NLSs and random peptides not overlapped with NLS. (b) NetSurfP RSA values of the 200 residues either side of the annotated NLSs and random peptides not overlapped with NLS. The index 0 represents the residue at the boundary of the left or right side of the NLS (or peptide).

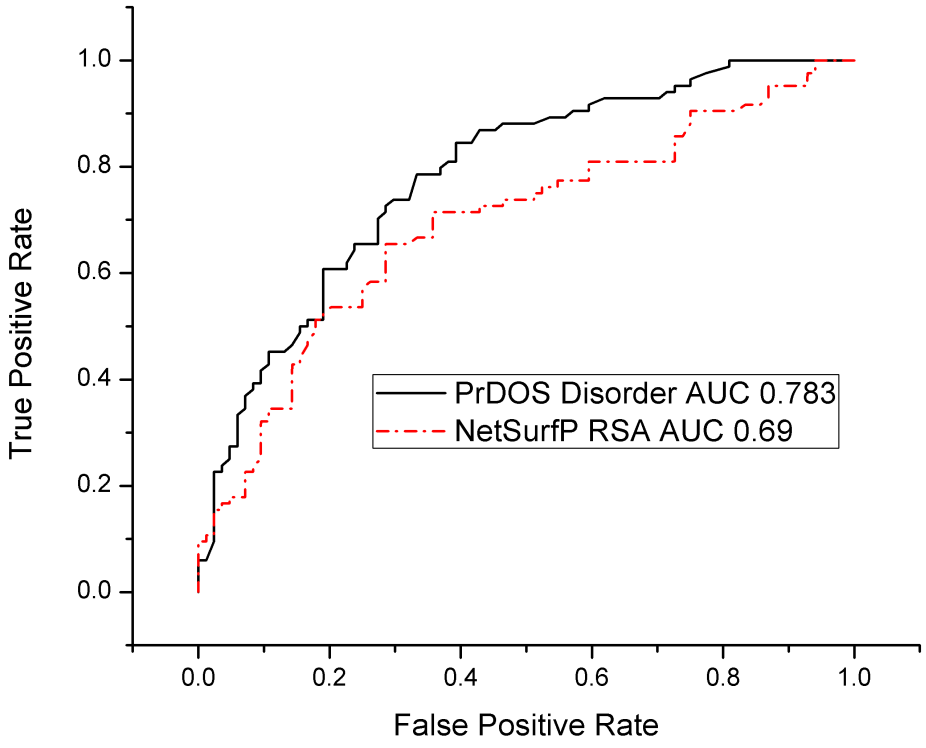


Figure 3.4: ROC curves for the PrDOS disorder feature and NetSurfP RSA feature.

between NLS and non-NLS peptides. We further used each of the attributes to build a single-feature SVM classifier. The LIBSVM package with the radial basis function kernel was used to run a 5-fold cross-validation on the 90 NLSs and 90 non-NLS peptides in the training dataset. We found that when the PrDOS disorder score of the peptide was used as the single feature, it achieved a 5-fold cross-validation accuracy of 70.83% on discriminating NLS and non-NLS peptides; while using the NetSurfP RSA score of the peptide as the single feature, it achieved 64.88% accuracy; when both the PrDOS disorder score and the NetSurfP RSA score of the peptide were used as the features, the accuracy was 70.24%, which was not higher than that of using the PrDOS disorder score alone. This indicates that although the NetSurfP RSA score is also a discriminative attribute, it is redundant if the PrDOS disorder score is used.

Therefore, in our following experiments only the PrDOS disorder score is used in the linear motif scoring to estimate the probability of being NLS.

3.3.3 Performance of the integrated predictor: SeqNLS

Figure 3.5 shows the prediction performance of SeqNLS on the Yeast and Hybrid datasets. The algorithm attains a precision and recall of 0.7 and 0.5 or higher when the final-score cutoff is set as 0.85. By tuning the final-score cutoff, the algorithm can attain different precision and recall rates with the higher final-score cutoff leading to higher precision and lower recall. The higher final-score cutoff also leads to the higher Mean aPC, which indicates that matches with the higher final scores generally are less likely to cover non-NLS amino acids. As indicated previously, the highest precisions of the sequence-based predictor are 0.667 and 0.77 respectively on the Yeast dataset and the Hybrid dataset by maximizing the enrichment-score cutoff. For the integrated predictor, precision can be further improved to around 0.75 to 0.8 on both the Yeast dataset and the Hybrid dataset while a higher recall is maintained. This implies that the proposed linear motif scoring and IRLC-masking improve the prediction. Figure 3.5 also shows that recall starts dropping dramatically when the final-score cutoff exceeds certain value over 0.8 on the both Yeast and Hybrid datasets. This is because matches with the enrichment scores higher than E_K certainly have the final scores at least 0.8 according to the formula of calculating the final score. These matches cover 56% and 60.3% of NLSs (see Figure 3.2(b) and Figure 3.2(d)) in the Yeast and Hybrid datasets. Therefore, recall won't drop dramatically when the final-score cutoff is lower than 0.8. When the final-score cutoff is set higher than 0.8, matches with the low enrichment scores are removed since the weight of the enrichment score is much higher than that of the linear motif score (0.8 vs. 0.2); with the increase of the cutoff afterward, matches with high enrichment scores but low linear motif scores will start being removed, and eventually only matches with

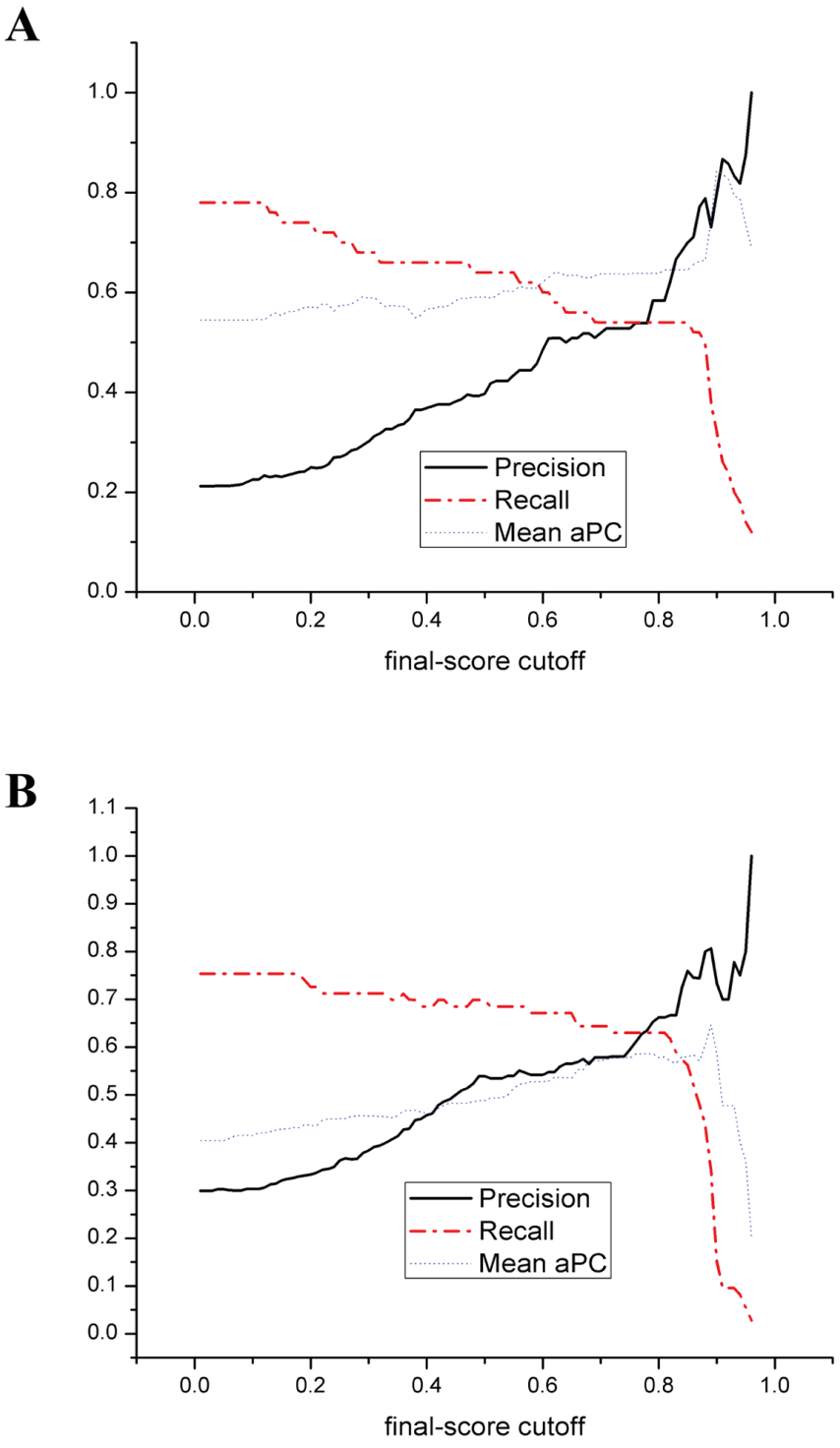


Figure 3.5: The prediction performance of the integrated predictor. (a) The Yeast dataset (b) The Hybrid dataset. IRLC masking is applied in both (a) and (b).

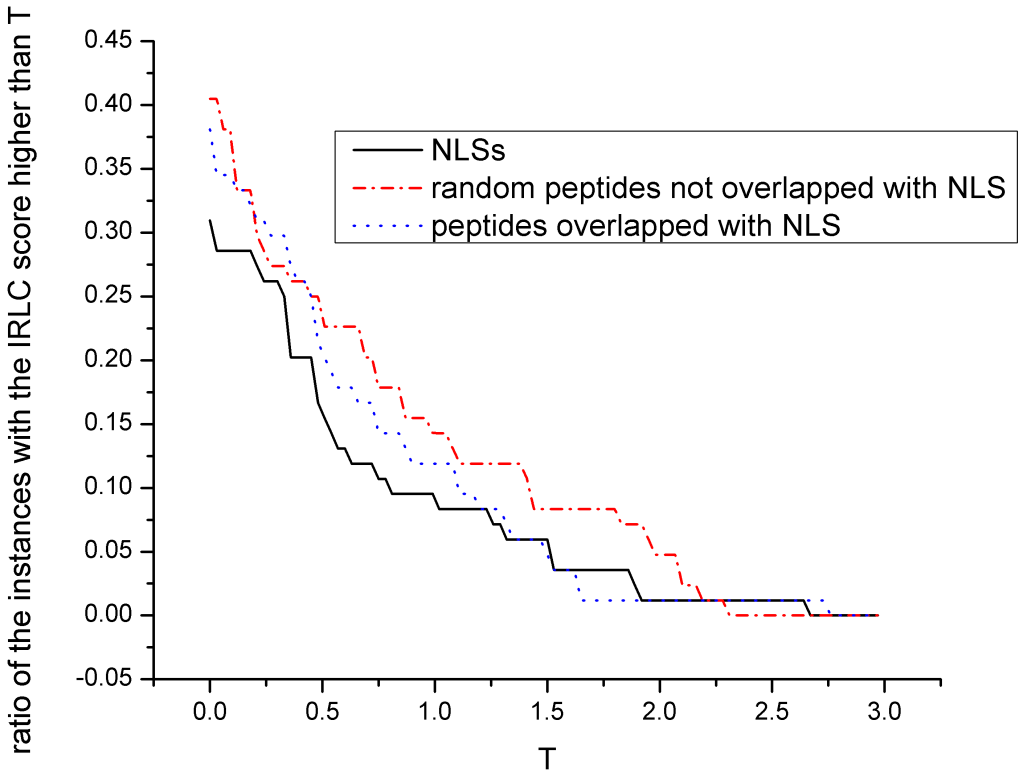


Figure 3.6: The effect of IRLC-masking. Peptides overlapped with NLS are obtained by adjusting boundaries of the NLSs to upstream or downstream proteins randomly in the parent by 1/3 length of the corresponding NLSs.

high enrichment scores and high linear motif scores are left. The result shows that for both the Yeast and Hybrid datasets, the precision of the integrated predictor can still be improved by increasing the final-score cutoff even when the final-score cutoff is already higher than 0.8. This indicates that matches with low linear motif scores are less likely to be (part of) NLS despite their high enrichment scores. Therefore, the enrichment score and the linear motif score are highly complementary in discerning NLS.

3.3.4 Effect of IRLC-masking

Figure 3.6 shows the ratio of three types of peptides in our training dataset with the IRLC scores higher than a threshold value T. It can be observed that the ratio of NLSs with the IRLC score higher than T is smaller than that of random peptides that are not overlapped with NLS. This result corresponds to our IRLC hypothesis that the chance is relatively low to find a residue in the flanking region of a NLS that is much more conserved; in other words, NLSs indeed tend to have higher relative local conservation. The similar trend can be observed for peptides partially overlapped with NLSs, which mimics true positive NLS predictions. This implies that IRLC-masking may be effective in masking out false positive NLS predictions with a smaller chance of masking out true positive NLS predictions. Figure 3.6 also shows that when T is higher than 1.7, both the ratios of NLSs and peptides overlapped with NLS with the IRLC score higher than T are close to 0. To avoid masking out any true positive predictions, the IRLC-masking cutoff is set as 1.7 throughout our experiment.

Table 3.3: The prediction performance of the integrated predictor with different final-score cutoffs with and without IRLC masking on the Yeast dataset

[1]Final		0.65	0.7	0.75	0.8	0.85	0.9	0.95
Precision		0.462	0.483	0.492	0.537	0.651	0.8	0.875
	IRLC	0.509	0.518	0.528	0.583	0.7	0.8	875
Recall		0.58	0.56	0.56	0.56	0.54	0.32	0.14
	IRLC	0.56	0.54	0.54	0.54	0.54	0.32	0.14
F1 scorel		0.514	0.519	0.524	0.548	0.59	0.457	0.241
	IRLC	0.533	0.529	0.534	0.561	0.61	0.457	0.241
Mean aPC		0.635	0.638	0.639	0.639	0.644	0.844	0.734
	IRLC	0.634	0.637	0.638	0.638	0.644	0.844	0.734

[1]Final-score cutoff

Table 3.3 and Table 3.4 describe the prediction performance of the integrated predictor with or without IRLC-masking on the Yeast and Hybrid datasets respectively. It shows that IRLC-masking improves the precision of the integrated predictor on the Yeast dataset while it is not effective on the Hybrid dataset. This is because the

Table 3.4: The prediction performance of the integrated predictor with different final-score cutoffs with and without IRLC masking on the Hybrid dataset

[1]Final		0.65	0.7	0.75	0.8	0.85	0.9	0.95
Precision		0.564	0.583	0.6	0.662	0.759	0.733	0.8
	IRLC	0.565	0.578	0.595	0.662	0.759	0.733	0.8
Recall		0.685	0.658	0.644	0.63	0.562	0.151	0.055
	IRLC	0.671	0.644	0.63	0.63	0.562	0.151	0.055
F1 scorel		0.619	0.618	0.621	0.646	0.646	0.25	0.103
	IRLC	0.614	0.609	0.612	0.646	0.646	0.25	0.103
Mean aPC		0.539	0.575	0.58	0.578	0.579	0.587	0.361
	IRLC	0.535	0.572	0.577	0.578	0.579	0.587	0.361

[1]Final-score cutoff

effect of IRLC-masking depends on where false positive predictions are distributed: if no false positive predictions are located in the regions of the sequence that contradict the property of relative local conservation (RLC), the precision cannot be improved. This can also explain why precision is not improved by applying IRLC-masking on the Yeast dataset when the final-score cutoff is higher than or equal to 0.9. In addition, it shows that for the both datasets after IRLC-masking is applied, recall decreases slightly when the final-score cutoff is lower than 0.8 while it remains the same when the final-score cutoff is higher than 0.8. This is because the true positive predictions coming from matches with the lower final scores generally have less accurate boundaries, which lead to more true positive predictions being masked out by IRLC-masking.

3.3.5 Comparison of SeqNLS with state-of-the-art NLS prediction algorithms

Here we compare the prediction performance of SeqNLS with those of state-of-the-art NLS prediction algorithms. Considering that some of the compared NLS predictors may generate overlapped NLS predictions, for all the compared NLS predictors, if two NLS predictions are overlapped, they will be merged into one prediction before evaluation. Table 3.5 and Table 3.6 show the prediction performance of different NLS

Table 3.5: The prediction performance of different NLS predictors on the Yeast dataset

Yeast Dataset	Precision	Recall	F1 score	Mean aPC
PSORT II	0.455	0.66	0.538	0.696
PredictNLS	0.462	0.12	0.19	0.411
NLStradamus	0.864	0.36	0.508	0.473
cNLS Mapper	0.8	0.46	0.584	0.437
NucImport	0.526	0.4	0.455	0.414
[1]Seq	0.569	0.56	0.564	0.641
[1]Int	0.7	0.54	0.61	0.644

[1]Sequence-based predictor (enrichment-score cutoff=1.62 (E_K))
[2]Integrated predictor (final-score cutoff = 0.85, IRLC masking)

Table 3.6: The prediction performance of different NLS predictors on the Hybrid dataset

Hybrid Dataset	Precision	Recall	F1 score	Mean aPC
PSORT II	0.617	0.671	0.643	0.657
PredictNLS	0.857	0.151	0.256	0.455
NLStradamus	0.714	0.329	0.45	0.56
cNLS Mapper	0.696	0.425	0.527	0.466
NucImport	0.632	0.329	0.432	0.358
[1]Seq	0.682	0.603	0.64	0.57
[1]Int	0.759	0.562	0.646	0.579

[1]Sequence-based predictor (enrichment-score cutoff=1.62 (E_K))
[2]Integrated predictor (final-score cutoff = 0.85, IRLC masking)

prediction methods on the Yeast and Hybrid datasets respectively. We can see that PSORTII has the highest recall on the both datasets while its precision is the lowest among all the methods. This indicates that many NLSs and non-NLS peptides can be covered by the NLS patterns used in PSORTII. An interesting observation is that PSORTII has the highest Mean aPC. We investigated the individual patterns used in PSORTII and found that its high Mean aPC is attributed to the predictions of the bipartite-NLS pattern (K/R)(K/R)X_{10}(K/R)$_{3/5}$. PredictNLS only generated a small number of predictions as shown by its low coverage in terms of recall. It was found that both NLStradamus and cNLS mapper have very high precision on the Yeast

dataset. This is partially due to that our Yeast dataset is included in the training data of the NLStradamus prediction server and the activity profiles built in cNLS mapper are optimized for yeast. For the Hybrid dataset, both NLStradamus and cNLS mapper exhibit lower precision since this dataset is not overlapped with the Yeast dataset and includes many different species in addition to the yeast species, of which the collected NLSs are from literature after 2010. The NucImport algorithm has a very poor Mean aPC score because its NLS predictions have uniform length of 20 amino acids. Another limitation of NucImport is that it can predict only one NLS per sequence while in the testing datasets some NLSs occur within the same parent proteins.

As shown in Table 3.5 and Table 3.6, our sequence-based predictor with the enrichment-score cutoff set as 1.62 (E_K) has comparable or better prediction performance than other NLS prediction methods: it achieved a recall rate of 0.56 and 0.603 on the Yeast dataset and the Hybrid dataset respectively, which is only second to PSORTII. However, its precision is better than PSORTII on both datasets. The integrated predictor shows better precision than the sequence-based predictor since it incorporates linear motif attributes. When the final-score cutoff is set as 0.85, the integrated predictor achieved a precision of 0.7 and 0.759 on the Yeast and the Hybrid datasets respectively while its recall is 0.54 on the Yeast dataset and 0.562 on the Hybrid dataset. That is, over 50% of the NLSs can be covered. The reason that the integrated predictor can achieve high precision while maintaining high recall is that the algorithm can extensively detect potential NLSs by using the sequential-pattern mining method while exploiting linear motif scoring, which is not used by other NLS prediction methods. As for residue-level accuracy, both the sequence-based predictor and the integrated predictor achieve the higher Mean aPC compared to most other NLS prediction methods because of its incorporation of the bipartite-NLS motif. It is interesting to note that another example of achieving better prediction performance

by integrating sequence features and predicted disorder is NESsential [104], which is a computational method designed to predict nuclear export signals (NESs).

3.4 Conclusions

In this study, we propose SeqNLS, a novel method for nuclear localization signal prediction based on frequent pattern mining and linear motif scoring. Various attributes of NLS including the sequential-pattern enrichment, predicted disorder, and local conservation are investigated based on the two well-curated datasets, which demonstrates their discriminative capacity for identifying NLSs. Our experimental results indicate that sequence features in terms of sequential patterns and linear motif features are highly complementary for NLS prediction. Compared to other state-of-the-art NLS prediction methods, SeqNLS achieves better overall prediction performance. For the Yeast and Hybrid datasets, SeqNLS attains a F1 score of 0.61 and 0.646 respectively compared to 0.538 and 0.643 of PSORT-II.

Chapter 4

Computational identification of post-translational modification (PTM) based nuclear import regulations by characterizing nuclear localization signal-import receptor interaction

4.1 Background

As the control center of the cell, the nucleus is separated from the cytoplasm by the nuclear envelop. Except that small nuclear proteins (<40 kDa) can diffuse into the nucleus through the nuclear pore complex (NPC) [105], most other nuclear proteins are imported into the nucleus through nuclear import pathways [106]. Traffic of large molecules such as proteins and RNA through the pores is required for both gene expression and the maintenance of chromosomes. Understanding of the regulation mechanisms of such traffic can bring biological insights to the cell and diseases. For example, deregulation of nuclear import is associated with numerous cancers such as breast cancers, prostate cancers, and other diseases [7–10].

In the nuclear import pathways, the import receptors, which belong to the karyopherins family, bind to NLS of the nuclear proteins. Proteins bound by import receptors can then pass through the NPC through the transient interaction between karyopherins-ß and NPC (2). NLS thus has been regarded as evidence for nuclear localization of proteins. Computational prediction of NLS has been a well-studied research topic [19–23] and current computational NLS prediction tools have substantially reduced time and cost for biologists to discover NLSs by focusing their

experiments on putative NLS motifs. However, nuclear import of proteins is more than an issue of binary outcomes (either nuclear localized or not nuclear localized). In fact, nuclear import is a highly regulated process [5, 6], in which the distribution of particular proteins between the nucleus and cytoplasm can be regulated through promoting or repressing their nuclear import activities.

The nucleus is the most critical compartment in a eukaryotic cell, which hosts important nuclear proteins such as transcription factors, histones, and signaling molecules. The import regulation of a particular nuclear protein is thus a means to control gene expression, cell proliferation, cell apoptosis, etc. in reaction to environmental changes [5, 7, 26, 27]. In recent years, an increasing number of researches are devoted to studying nuclear import regulation of proteins. The discovery of the import regulation mechanism for a particular nuclear protein is of great interest since it implies a potential way to control the protein's activity [5]. Moreover, it contributes to uncovering the potential biological pathways that regulate the associated biological activities in the nucleus. Nuclear import activity is mostly regulated through modulating the interactions between nuclear proteins and their binding import receptors [8]. In particular, modulating the NLS binding affinity to its binding receptor by post-translational modification (PTM) is the best understood mechanism (PTM-based nuclear import regulation) that regulates the nuclear import of proteins. In previous studies, the most common type of PTM for nuclear import regulation is phosphorylation [5, 6, 26–28] while lysine acetylation has been found to be another frequent type [29–36]. The reason that nuclear import can be regulated through the PTM is that nuclear import activity is directly related to the binding affinity of NLS for its binding import receptor [37–39]. However, identifying the PTM-based nuclear import regulation is difficult since PTM may promote, repress or may not have obvious impact on the nuclear import activity [27].

The most commonly used approach to infer the PTM-based nuclear import regu-

lation is the site-directed mutagenic analysis [26, 32, 40–46]. This approach basically mutates the NLS residue at the PTM site to a residue that either prevents the PTM or mimics the residue after the PTM. It then evaluates the likelihood that the PTM regulates the nuclear import of the protein based on the change of the corresponding nuclear import activity. The strategy of mimicking residue after PTM such as phosphorylation has been performed computationally by cNLS mapper [22], in which the position-wise contributions of different amino acids to the nuclear import activity are approximated in the activity-based profiles. However, the interaction between a NLS and its binding import receptor is very sensitive to the NLS change. The site-directed mutagenic analysis is thus not always reliable due to the difference between the mimicking residues and the residues after PTM. Since the PTM-based nuclear import regulation is now recognized as a common nuclear import regulation mechanism, there is a need for developing quantitative methods to expand the identification of more PTM-regulated nuclear proteins [27].

For the PTM-based nuclear import regulation, it is technically true that PTM regulates the nuclear import of a protein through modification of its NLS residue(s). However, the induced change on the interaction between the NLS and the import receptor is the ultimate factor that governs the change on its nuclear import activity. In other words, the induced change on the NLS-import receptor interaction should better characterize the change of the nuclear import activity caused by PTM than the difference of the NLSs. Therefore, in our method we first applied molecular interaction energy components (MIEC) [47–50], which has been successfully used to characterize domain-peptide interactions, to characterize the NLS-import receptor interaction. Next, we used SVR to learn the relationship between the MIEC features and the corresponding nuclear import activity, which is quantitated as NLS activity scores [22] in the experimental dataset. The characteristic of our method (NIpredict) is that it is a machine learning based method based on features calculated from

NLS-import receptor interaction interface, which can thus be applied to assess the impact of PTM within NLS on the corresponding nuclear import activity. Our cross-validation results showed that nuclear import activities for different NLS variations can be accurately predicted by NIpredict. We then applied NIpredict systematically to identify the potential PTM-based nuclear import regulations for human and yeast nuclear proteins.

4.2 Material AND Methods

4.2.1 Preparation of the training dataset

The training datasets of NIpredict were prepared based on the experimental dataset from Kosugi et al [22]. Kosugi's dataset provides the NLS activity scores as the measure to represent different levels of nuclear import activities for a number of classical NLS mutants. Considering that the major binding site and the minor binding site in Imp-α are different binding site and may have different interactions with the bound NLSs, two training datasets for NIpredict were prepared as shown in the supplementary file (Table S3 and Table S4): the major-site dataset is for NLSs bound to the major binding site with 374 instances while the minor-site dataset is for NLSs bound to the minor binding site with 152 instances In the major binding site, there are five well-recognized binding site positions (P1-P5), while there are four well-recognized binding site positions (P1'-P4') in the minor binding site [70, 91, 107]. The alignment of the NLS residues in the dataset onto the binding site positions can be obtained by aligning the strictly conserved lysine in P2 and the KR-motif in P1'P2' [70, 107] for the major-site and minor-site datasets respectively. Since the experiment of Kosugi's dataset was conducted in vivo in living yeast cells, the import receptor was implicitly indicated as Kap60 since Saccharomyces cerevisiae possesses a single Imp–α gene.

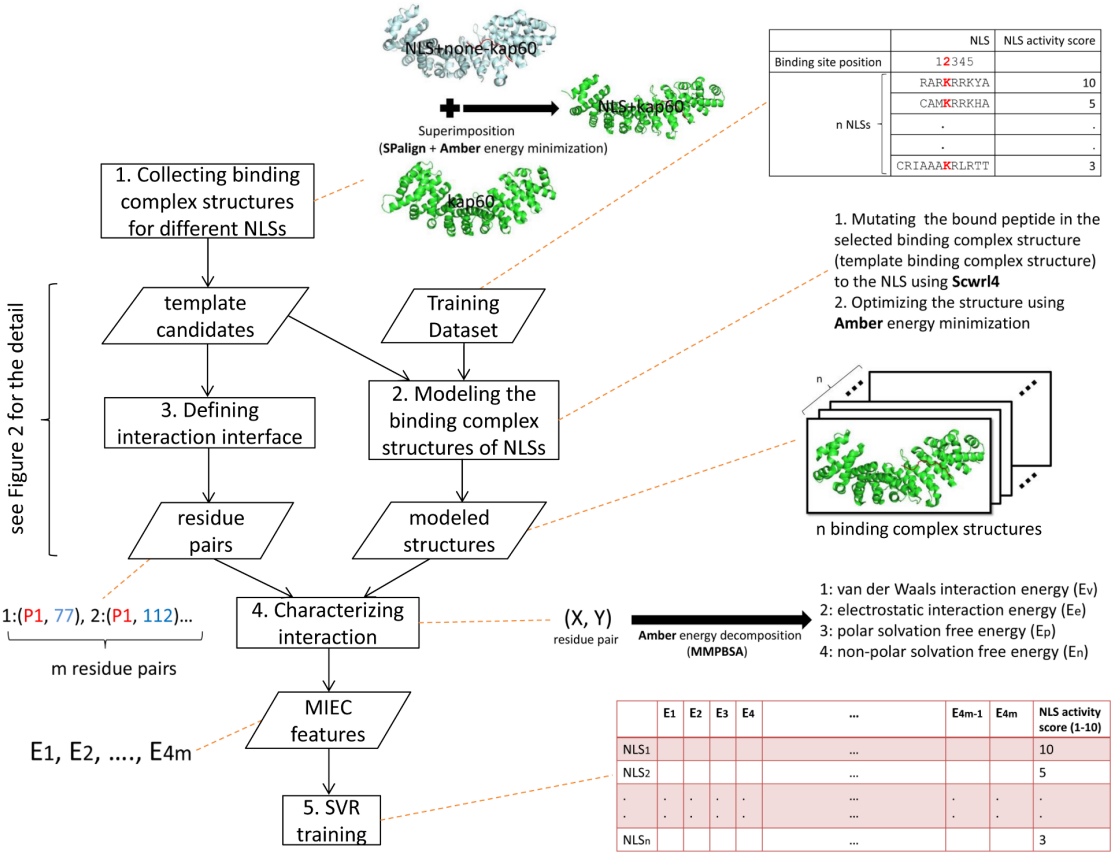


Figure 4.1: Flowchart of building the NIpredict model. The binding complex structures of different NLSs were first modeled by the defined superimposition procedure (step 1). The binding complex structure of NLS mutants were modeled by mutating the NLS residues in the selected template using Scwrl4 and then optimizing the structure through the defined Amber energy minimization procedure (step 2). The interaction interface was modeled in terms of a set of residue pairs (How to derive the set of residue pairs in step 3 will be explained in Figure 2). For each of the modeled binding complex structures, the interaction between the NLS and the import receptor was characterized in terms of the four decomposed energies between each of the m residue pairs (step4). Support Vector Regression was used learn the relationship between the characterized interaction, which is represented by MIEC features, and the corresponding NLS activity score (step5).

4.2.2 Building the NIpredict model

The process of building the NIpredict model is shown in a flowchart (Figure 4.1). First, we collected binding complex structures for different NLSs. In step 2, we modeled the binding complex structures of the training dataset (see next section). Next, we defined NLS-import receptor interaction interface. In step 4, the NLS-import receptor interactions of the training dataset are characterized based on MIEC and Amber simulation. Finally, the relationship between the characterized interaction and the corresponding nuclear import activity is learned by a Support Vector Regression model. The details of each step will be discussed in the following subsections.

Collecting binding complex structures for different NLSs

We collected a number of NLS-Impa binding complex structures from the PDB database [108] with different NLSs as template candidates. This is because different NLSs may form different orientations when bound to kap60, and we collected template candidates for two reasons: One is to construct the generic model of the NLS-import receptor interaction interface; the other is to choose an optimal structure template to model the binding complex structures of NLSs in our training datasets.11 NLS-Imp binding complexes were collected for the major binding site: 1EE4, 1IQ1, 1Q1S, 1Y2A, 2YNR, 3BTR, 3OQS, 3RZ9, 3VE6, 4BA3, and 4HTV; 5 NLS-Imp binding complexes were collected for the minor binding site: 1EE4, 1IQ1, 1Q1S, 2YNR, and 3Q5U. However, among those collected complexes most of the binding proteins were not kap60 except 1EE4. To expand the pool of template candidates, for each of the collected binding complex structures of which the binding protein was not kap60 (binding complex structure C), we built the corresponding binding complex structure of the bound NLS peptide in C for the binding import receptor as kap60 (binding complex structure C') through the defined superimposition procedure: C' was built by first superimposing C to the yeast NLS binding complex 1EE4

and then acquiring the coordinates of the bound peptide, which was followed by the energy-minimization procedure (to be defined below) to optimize the structure. The structure superimposition was performed by SPalign [109].

Modeling binding complex structures

The procedure to model the binding complex structure of a NLS A for its binding import receptor can be divided into two steps: 1) selecting a template binding complex and mutating its bound peptides to the NLS A; 2) optimizing the structure. In the first step, the bound peptides in the complex were truncated to exactly align the binding site positions of the NLS A to avoid the influence from the extra flanking residues while SCWRL4 [110] was used to mutate the truncated bound peptides to the NLS A In the second step, the mutated complex structure from SCWRL4 was optimized by performing energy minimization using AMBER12 and the AMBER03 force field [111] which followed the same procedures by Li et al [50]: before performing energy minimization, tleap was first used to preprocess the complex structure so that the structure was solvated in a rectangle box of TIP3P water that extended 12 Å from any solute atom, and the system was neutralized by placing counter ions Na+ or Cl- around the structure using the Columbia potential. The preprocessed complex structure was then optimized in 5000 steps of energy minimization, in which the first 1000 steps used steepest descent minimization and the rest 4000 steps used conjugate gradient minimization. The snapshot of the conformation in the last step was used for energy decomposition (to be explained below).

The template binding complex determines the starting conformation of the modeled complex structure and thus affects the conformation of the optimized binding complex structure. Selection of the template binding complex is therefore highly associated with the accuracy of the modeled binding complex structure. The template binding complex was selected from the template candidates according to the corre-

sponding performance, while the template candidates of which the bound peptides did not completely cover the binding site positions of the NLS were not considered. 3RZ9 was used as the template complex structure for NLSs in the major-site dataset except NLSs numbered from 306 to 374; the binding sites positions of those NLSs numbered from 306 to 374 cannot be completely covered by the bound peptide in 3RZ9 and 1Q1S was used as their template binding complex structure. For the minor-site dataset, no bound peptides in the collected PDB structures were found to completely cover the binding sites positions of the NLSs. 3UKX was used as the template complex structure since its bound peptide covers the most binding site positions of the NLSs in the minor-site dataset (except P-4' for NLS numbered from 80 to 152 in the minor-site dataset). To model the position of the missing NLS residue serine in P-4'position, residue serine was placed in P-4'position of 3UKX using Swiss-PdbViewer [112].

Modeling interaction

The basic idea of modeling interaction is to identify the residue pairs between NLS residues and/or import receptor residues that may affect binding and then characterize their preference using their interacting energies. Three classes of residue pairs were defined which are composed of residues at two positions of NLS and/or kap60p. The most important class is the domain-peptide residue pair, which was defined as residue pairs between the directly interacted NLS residues at the binding site positions P1-P5 (P1'-P4' for the minor binding site) and residues in kap60 within a 6-Å distance cutoff for any of the template candidates. Accordingly, 104 domain-peptide residue pairs were defined for the characterization of the domain-peptide interaction for the major site, which were highlighted by the orange links in Figure 4.2(a) and listed in Figure 4.2(b); the potentially interacting residues in kap60 are highlighted in blue color in Figure 4.2(c). For the minor site 80 domain-peptide residue pairs were defined. In addition, the interactions between the directly interacted NLS

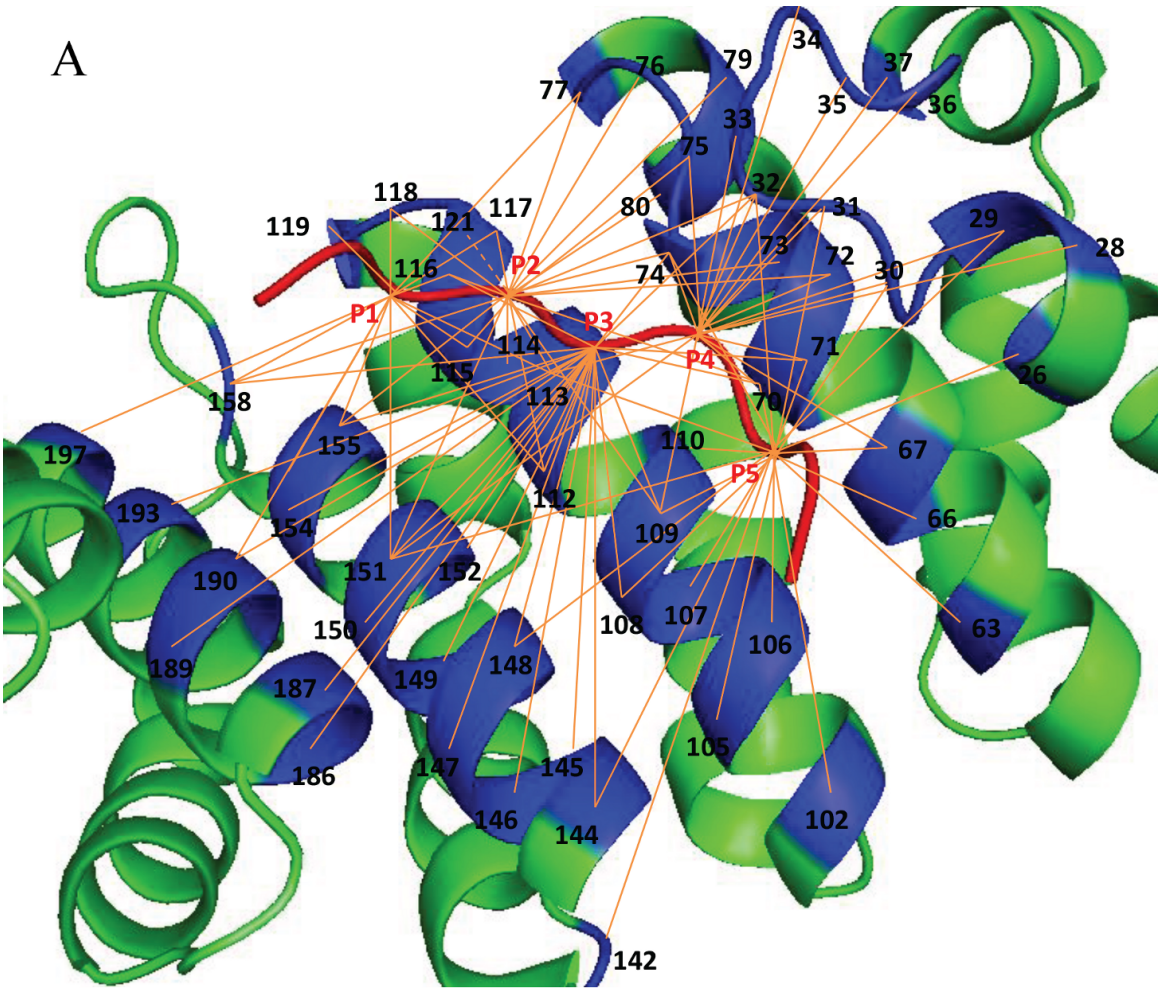


Figure 4.2: Residue pairs defined at the domain-peptide interaction interface for the major site. (a) Diagram of the modeled interaction interface. Orange links represent the domain-peptide residue pairs. (b) List of the defined domain-peptide residue pairs. (c) Potentially interacting residues in kap60.

residues in the adjacent positions (internal-peptide interaction) were also considered to reflect the conformational preference of the peptide under the bound condition: 4 and 3 internal-peptide residue pairs were defined for the characterization of the internal-peptide interaction for the major site and the minor site respectively. In addition, interactions of NLS flanking residues were considered (flanking-residue interaction) to reflect their enhancing or repressing effect on NLS binding: 2 N-terminal flanking residues and 2 C-terminal flanking residues next to the directly interacted NLS residues were defined as the NLS flanking residues. Residue pairs between the NLS flanking residues and the directly interacted residues along with their previously defined potentially interacting import receptor residues were defined as the flanking-residue residue pairs, which were used for the characterization of the flanking-residue interaction: 256 and 220 flanking-residue residue pairs were thus defined for the major site and the minor site respectively. The combination of domain-peptide residue pairs, internal-peptide residue pairs, and flanking-residue residue pairs was used to define the interaction interface.

Given the modeled binding complex structure (the snapshot of the conformation in the last step of the energy minimization procedure) and the defined residue pairs of the interaction interface, the interaction can be characterized in terms of the decomposed energies between the residue-pairs defined in the model of the interaction interface. Energy decomposition was performed using MMPBSA of the AMBER package, while the GB model was used to simulate the solvent effect with the input parameters developed by Onufriev et al [113, 114]. The interaction energies between any two residues of interest were decomposed into van der Waals interaction energy, electrostatic interaction energy, polar solvation free energy, and non-polar solvation free energy. The decomposed energies between the defined residue pairs in the model of the interaction interface were used as the input features of the SVR, while the NLS activity scores were used as the prediction values. The SVR model was trained using

the LIBSVM package [100] with nu-support vector regression (nu-SVR) as the SVM type and linear function as the kernel.

4.2.3 Performance Evaluation

To evaluate the prediction performance of NIpredict, we performed leave-one-out cross-validation on the training datasets. Two metrics were used to evaluate the prediction performance with regard to the NLS activity scores: correlation coefficient and mean squared error (MSE). Since the training datasets are composed of activity scores of different variations of the NLS signal, the performance of NIpredict on the training datasets can be regarded as evaluating how accurate NIpredict can assess the impact of NLS change over its nuclear import activity. Direct evaluation on how effective NIpredict can identify the PTM-based nuclear import regulation will be discussed in the case studies.

4.2.4 Genome-wide prediction of nuclear import activity

In addition to performance evaluation on the training datasets, we applied NIpredict to predict the nuclear import activity of genome-wide nuclear proteins that contain targeted peptides covered by the mutation templates (shown in Table S3) in our training datasets. This limited coverage is due to that our prediction model is trained on NLS instances from these templates. Accurate prediction of binding activity of proteins of other templates depends on the experimental data of their binding activities, which are not available now. While most mutation templates in the training datasets are not common in proteins of human and yeast genomes, the mutation template PxxK[KR]x[KR]xx is a very common NLS motif bound to the major site. In this motif, the P2 site lysine is strictly conserved; P3 and P5 are conserved for basic residues either K or R, and a helix breaking proline is located in the N terminal flanking position P-2. We prepared the Yeast dataset which contains 1404 nuclear proteins

collected from Uniprot. Besides, although these training datasets are for the yeast species and the import receptor is the unique kap60, previous study showed that the NLSs bound to the major site or minor site of the Imp-α in yeast are also bound to most Imp-α variants in human [87]. This implies that the binding specificities of the Imp-α variants in human are similar to kap60's. Therefore, in addition to the Yeast dataset, we prepared the Human dataset which contains 2720 nuclear proteins from Uniprot. We scanned the Yeast and Human datasets for matches of the motif PxxK[KR]x[KR]xx and then applied NIpredict to predict the import activity of these matched proteins. To model the NLS modified by PTM, the Amber library file of the phosphorylated amino acids based on craft et al [115] and the Amber library file of the acetylated lysine based on papamokos et al [116] were used.

4.3 Result AND Discussion

4.3.1 Characterization of NLS-Impa interaction using MIECs

In NIpredict, the decomposed interaction energies of all the residue pairs defined in the model of the interaction interface are combined as the features for characterizing the interaction between NLSs and import receptors. To measure the energy contribution of a peptide residue to the interaction, all the decomposed energies between this peptide residue and each of its potentially interacting domain residues as defined in the domain-residue residue pairs are added up. Figure 4.3 shows the average energy contributions of the binding site positions P1-P5 and P1'-P4' for the major-site dataset and the minor-site dataset respectively. For the major site, the figure shows that position 2 (P2) has the highest average energy contribution, while energy contributions of P3 and P5 are the second and roughly equal contribution, which is consistent with the results of previous research [117]. For the minor site, it is shown that P2' has the highest energy contribution while P1' is second to P2'. The importance of these two positions in terms of energy contribution corresponds to the

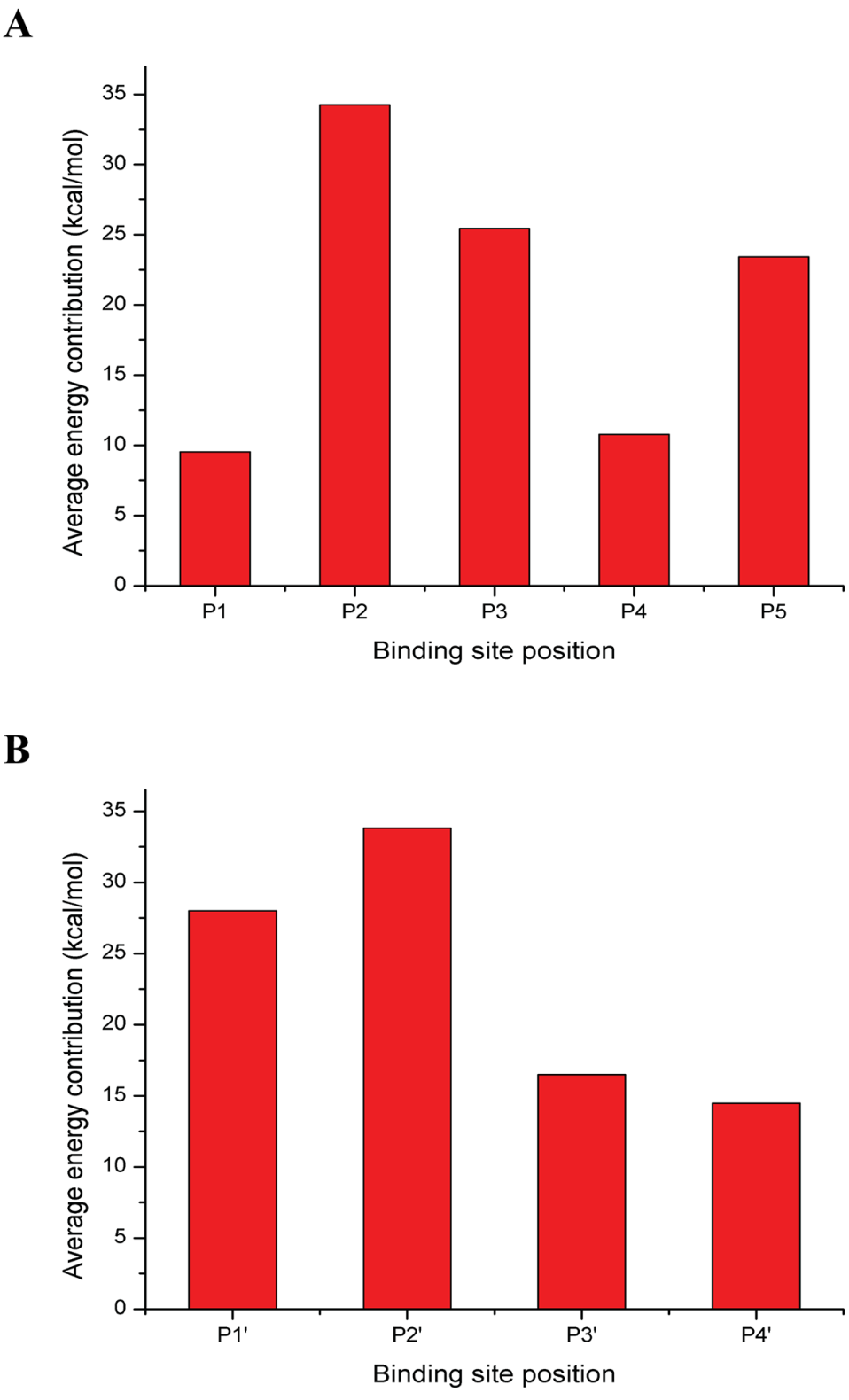


Figure 4.3: The average energy contributions of different binding site positions. (A) The major binding site; (B) The minor binding site.

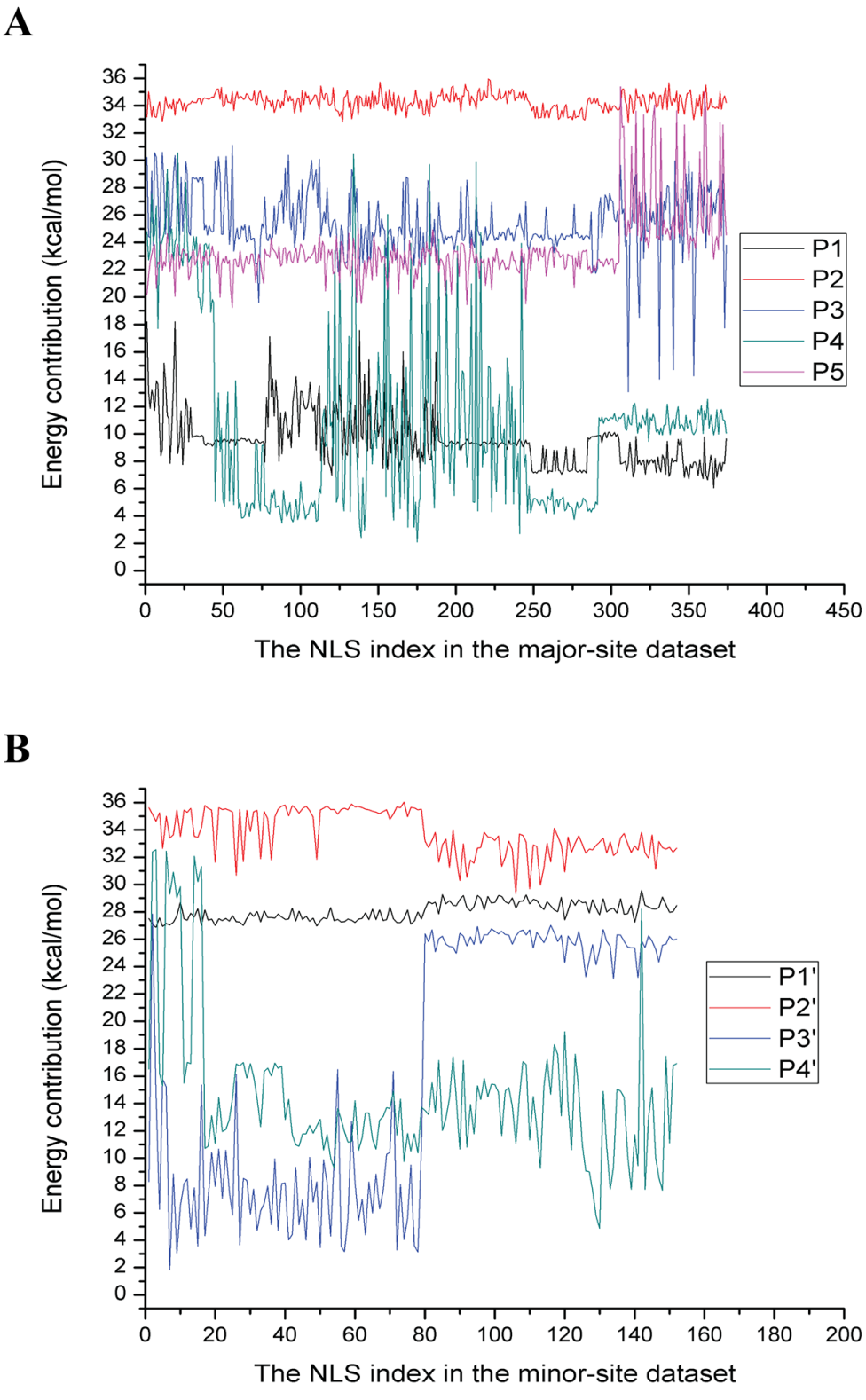


Figure 4.4: The energy contributions of different binding site positions for NLSs in the datasets. (A) The major binding site; (B) The minor binding site.

conserved "KR" motif at P1'P2', of which P2' arginine is especially important to the interaction [107]. The consistency between our predicted contribution of the peptides to the binding and previous studies demonstrates the effectiveness of our binding complex structure modeling the characterization based on decomposed energies. On the other hand, we found that the relative energy contributions for different binding site positions are not always the same as shown in Figure 4.4. They depend on the types of residues at the positions and the flanking residues, which implies the high flexibility of the binding site to affect NLS binding.

Table 4.1: Performance of NIpredict using different models of the interaction interface

	Major Site		Minor Site	
interaction interface model (residue pairs)	Correlation Coefficient	MSE	Correlation Coefficient	MSE
domain-peptide	0.637	4.16	0.773	3.47
domain-peptide+internal-peptide	0.637	4.15	0.771	3.5
domain-peptide+flanking-residue	0.729	3.45	0.713	4.61
domain-peptide+internal-peptide +flanking-residue	0.718	3.59	0.719	4.52

To illustrate how the definition of the interaction interface affects the binding activity prediction, Table 4.1 shows the performance of NIpredict using different interaction interface definitions. NIpredict achieved a correlation coefficient of 0.729 for the major-site dataset when the domain-peptide residue pairs and flanking-residue residue pairs were included in the interaction interface, while it attained a correlation coefficient of 0.773 for the minor-site dataset when only the domain-peptide residue pairs were used. From Table 4.1, we found that inclusion of the internal-peptide residue pairs had no performance improvement for both the major-site dataset and minor-site dataset. The reason could be that NLSs bound to the major site or the minor site are short and no specific secondary structures are formed under the bound condition, which is also suggested before [50]. On the other hand, the inclusion of flanking-residue residue pairs significantly improved the performance on the major-

site dataset. One of the reasons could be that many NLSs in the major-site dataset are only different in the flanking residues, and the interaction with those flanking residues are thus more effective features for predicting the import activity of those NLSs. For the minor-site dataset, however, the inclusion of the flanking-residue residue pairs degraded the performance, which could be due to that flanking-residue residue pairs introduced too many irrelevant residue interactions into the interface model for the minor-site dataset. In machine learning it is well known that noise features downgrade SVR performance.

We also evaluated the leave-one-out prediction performance on the 76 NLSs of the major-site training dataset that are covered by the mutation template PxxK[KR]x[KR]xx in the major-site dataset. It achieved a correlation coefficient of 0.633 and MSE of 5.76. The reason of having lower prediction performance on those 76 NLSs than the overall prediction performance on the major-site dataset could be that the mutation template PxxK[KR]x[KR]xx includes more wildcard positions and involves more complicated variations than other mutation templates.

4.3.2 Comparison with other methods on predicting nuclear import activity

Both NIpredictand cNLS mapper can be used to predict nuclear import activity. NIpredict achieved a correlation coefficient and MSE of 0.729 and 0.345 on the major-site dataset and 0.773 and 0.347 on the minor-site dataset. cNLS mapper obtained a correlation coefficient and MSE of 0.881 and 1.782 on the major-site dataset and 0.871 and 1.468 on the minor-site dataset, of which the result was obtained by testing the datasets on the cNLS mapper website predictor. However, because the activity profiles of cNLS mapper were built based on our test datasets, its performance tend to be overestimated. Indeed, when the training and test datasets are the same as we tested cNLS mapper, NIpredict attained a correlation coefficient and MSE of 0.948 and 0.731 on the major-site dataset and 0.944 and 0.923 on the minor-site dataset,

which beats cNLS mapper.

In addition to predicting nuclear import activities for PTM-modified NLSs, NIpredict is expected to be more accurate on predicting nuclear import activities for wild-type NLSs than cNLS mapper. The reason is that while in general higher binding affinity leads to higher nuclear import activity, the exact relationship between binding affinity and corresponding nuclear import activity is complicated and is still not well understood [38]. So the additive rules incorporated in the activity profiles of cNLS mapper may be insufficient to model the complicated relationship between NLSs and their corresponding import activity. On the other hand, MIEC features have been successfully used to characterize various domain-peptide interactions. In particular, it was successfully applied to predict the binding affinity of another domain, Amphiphysin-1 Src homology 3, for different bound peptides [47]. This implies that the impact of bound peptide change on its binding affinity can be effectively characterized using MIEC, which is also proved by our successful prediction of nuclear import activity based on MIEC. In short, the advantage of NIpredict on predicting the nuclear import activity of wild-type NLSs lies that NIpredict is more sensitive to subtle changes within NLS that affect nuclear import activity.

4.3.3 Computational prediction of phosphorylation-based nuclear import inhibition of UL44

To further evaluate the capability of NIpredict in identifying PTM-based nuclear import regulation, we applied it to analyzing the regulation of the human virus UL44. It is known that the segment 425-PNTKKQK-431 of UL44 is a NLS bound to Impa/ß complex with high affinity and is sufficient and necessary for its nuclear localization [118]. Further studies found that the Thr-427 of the above NLS is a CDK1-mediated phosphorylation site which is related to its nuclear import. The phosphorylation at Thr-427 appears to inhibit its nuclear import based on the fact

that the mutation of Thr-427 to Ala-427 leaded to similar accumulation in the nucleus while Thr-427 to Asp-427 mutation, which mimics the phosphorylation, greatly reduced its nuclear activity [45, 119]. So it remains to explain how this is achieved. We found that the segment 425-PNTKKQKCG-433 NLS is covered by our template PxxK[KR]x[KR]xx while Thr-427 is at the wildcard position. We predicted the nuclear import activity of UL44 before and after Thr-427 phosphorylation using NIpredict. The predicted NLS activity score of the wild-type NLS PNTKKQKCG was 8.3 while the predicted NLS activity scores of the NLS mutants PNAKKQKCG and PNDKKQKCG were 7.7 and 4.9 respectively. Not unexpected, the predicted NLS activity score of the Thr-427 phosphorylated NLS was only 2.5, which indicates that the nuclear import is significantly inhibited due to phosphorylation. This prediction result by NIpredict was consistent with previous experimental studies [118]. Essentially, it computationally confirmed that Thr-427 phosphorylation inhibits the nuclear import of UL44 by weakening its interaction with Imp-α.

4.3.4 Genome-wide scan of PTM-based nuclear import regulations on the Yeast and Human datasets

To identify more potential PTM-regulated nuclear import proteins, we developed a systematic pipeline based on NIpredict to scan the Yeast and Human datasets. The first step is to identify potential NLSs by pattern match with motif PxxK[KR]x[KR]xx and then predict their nuclear import activity using NIpredict. The scan result is shown in Table S5 and Table S6 for the Yeast and Human datasets respectively. 117 motif matches were found for the Yeast dataset, of which 102 motif matches have predicted NLS activity scores greater than 3. 385 motif matches were found for the Human dataset, of which 331 motif matches have predicted NLS activity scores greater than 3. The motif matches that are overlapped with the Uniprot-annotated NLS (potential, portable, etc.) are marked as (*). In the Yeast dataset, 5 motif matches

are overlapped with the Uniprot annotated NLS while in the Human dataset, 34 motif matches are overlapped with the Uniprot annotated NLS. The lowest predicted NLS activity score of the motif matches overlapped with the Uniprot annotated NLS is 3.71, which implies that motif matches with very low predicted NLS activity scores are unlikely to be NLS.

In addition to the predicted NLS activity scores from NIpredict, another important factor to evaluate the likelihood of a candidate NLS is the disorder degree of the peptide segment. NLS is one type of linear motifs which have tendency to be located in disordered region [52]. The disorder factor was excluded from Kosugi's experiment since all NLS mutants were fused to the C-terminal of GUS-GFP reporter. However, some of the motif matches in the datasets may be located in globular domain where they may not be exposed to interact. Therefore, we estimated the disorder score of each motif match, which is defined as the medium disorder score of residues within the match. The disorder score of each residue was predicted using IUPred [120]. As shown in Table S5 and Table S6, all the motif matches (except one) that are overlapped with annotated NLS have disorder scores higher than 0.5. This indicates that motif matches with the disorder scores lower than 0.5 are less likely to be NLS. Therefore, the motif matches with the predicted NLS activity score higher than 3 and the disordered score higher than 0.5 were identified as potential NLSs for the further investigations.

To identify potential PTM-based nuclear import regulation, we collected the experimentally verified PTM sites from DbPTM 3.0 [121] that overlap the wildcard positions of the motif matches. The NLS activity scores of the motif matches before and after PTM were predicted using NIpredict and are shown in Table S7. The predicted NLS activity scores of the motif matches increase, decrease, or remain roughly equal after phosphorylation or acetylation on the underlined residue(s). All possible combinations of the PTM sites within the predicted NLS were listed in Table S7, while

some of the combinations may never happen during the life cycle of the protein. The motif matches of which the predicted NLS activity scores subject to apparent change are likely to be the candidates of the PTM-based nuclear import regulation, while the criteria of identifying potential NLSs were also considered to judge the likelihood of the PTM-based nuclear import regulation.

4.3.5 Case Studies on PTM-based nuclear import regulations identified by NIpredict

Localization of nuclear proteins to nucleus is the prerequisite for their participation in nuclear activities. Regulation of nuclear localization is a known mechanism to control and regulate other biological activities such as gene transcription and cell cycle progression [5]. It is thus interesting to check what biological activities could be regulated by our identified potential PTM-based nuclear import regulations. By screening the candidates of potential PTM-based nuclear import regulations in Table S7, we identified three potential regulation mechanisms of biological activities based on NIpredict predictions and associated biological evidences.

PKG inhibits spliceosome assembly by phosphorylation based regulation of SF1 nuclear import

A spliceosome is a large complex in the nucleus whose function is to remove introns from pre-mRNA (RNA splicing). Spliceosome assembly is thus a necessary event for RNA-splicing. SF1 and U2AF65 are both important components in spliceosome and their interaction is critical for spliceosome assembly. It was found that the Ser-20 phosphorylation mediated by cGMP-dependent protein kinase (PKG) on human SF1 (Q15637) inhibits its interaction with U2AF65, which leads to a block of spliceosome assembly [122]. Wang et al [122] raised another possibility that the Ser-20 phosphorylation may regulate localization of SF1 since Ser-20 is adjacent to a putative NLS, which is among our NIpredict predictions (Table S5). The motif match 13-

PSKKRKRSR-21 of SF1 has a predicted NLS activity score and a disorder score of 8.04 and 0.65, which is very likely to be NLS. As shown in Table S7, the predicted NLS activity score of this putative NLS decreased from 8.04 to 3.78 after the Ser-20 phosphorylation. The significant reduction on the predicted NLS activity score of SF1 indicates the repressing effect of the Ser-20 phosphorylation on the nuclear import of SF1. This result is consistent with the biological role of PKG with regard to SF1, which is known to prevent spliceosome assembly through repressing the nuclear activity of SF1 by the Ser-20 phosphorylation. Therefore, the PTM-based nuclear import regulation of SF1 mediated by PKG could be another mechanism to regulate RNA splicing.

SIRT1 promotes histone H1 nuclear import by decetylation

Histone is a protein family of which the function is to package DNA into structural units called nucleosomes. Formation of nucleosomes directly contributes to condensed chromatin, of which the repressive chromatin structure leads to DNA substrates less accessible for gene transcription. Histone acetylation/deacetylation is closely associated with gene transcription in that histone acetylation facilitates gene transcription through dissociating DNA from nucleosomes and de-condensing chromatins while histone deacetylation represses gene transcription through recovering nucleosomes and condensing chromatin. SIRT1 is a known histone deacetylase which deacetylates histone H1 (P10412) at Lys-26 [123] and catalyzes the formation of compacted chromatin. From out NIpredict predictions (Table S5), histone H1 has a motif match 19-PVKKKARKS-27, which has a predicted NLS activity score and a disorder score of 12.8 and 0.64. It means that this motif match is likely to be an NLS. As shown in Table S7, the Lys-26 acetylation on histone H1 reduced its predicted NLS activity score from 12.8 to 6.3, which indicates the promoting effect of the Lys-26 deacetylation on the nuclear import of histone H1. Considering that histone H1 is the prerequisite

component for organizing nucleosomes, increasing its nuclear availability could thus facilitate condensing chromatin. This is consistent with the biological role of SIRT1 with regard to histone H1, which is known to promote chromatin compacting through the Lys-26 decetylation on histone H1. Interestingly, the predicted NLS activity score before the Lys-26 deacetylation is still within the functional range (6.3), indicating that the basic level of nuclear availability of histone H1 is still required when gene transcription is active. These evidences showed that the PTM-based nuclear import regulation of histone H1 mediated by SIRT1 could be another mechanism to regulate gene transcription.

CDK regulates the nuclear import of ORC6 via phosphorylation

DNA replication occurs only in dividing eukaryotic cells, which must be tightly controlled to ensure that the genome is only replicated once. Previous studies have found that DNA replication is regulated by Cyclin-dependent kinases (CDK)-mediated phosphorylation on different proteins through multiple levels of mechanisms [124]. In particular, it is known that CDK mediated phosphorylation on ORC6 (P38826) has an effect to prevent helicase from loading with unknown mechanisms [124–126] One possible mechanism is that CDK mediated phosphorylation on ORC6 blocks Cdt1 recruitment through inhibiting Cdt1 binding [127]. We found that ORC6 contains a motif match 115-PSPKKNKRS-123 which is covered by our NIpredict predictions (Table S6). The predicted NLS activity score and the disorder score of this motif match is 5.04 and 0.67 respectively. This motif match is thus likely to be an NLS, in which Ser-116 is the phosphorylation site mediated by CDK [128] . As shown in Table S7, the predicted NLS activity score of this motif match drops from 5.04 to 1.27 after the Ser-116 phosphorylation, which indicates that the Ser-116 phosphorylation significantly inhibits the nuclear import of ORC6. This prediction result is consistent with the biological role of CDK with regard to ORC6, which is known to

prevent helicase from loading through CDK-mediated phosphorylation. Therefore, the PTM-based nuclear import regulation of the ORC6 mediated by CDK could be another mechanism to regulate DNA replication.

The above three hypotheses on PTM-based nuclear import regulation mechanisms were based on the prediction results of NIpredict and the known biological roles of PTM-mediating enzymes (PKG, SIRT1, and CDK). We found that a common characteristic of these three proteins is that their participations in the nuclear activities are also controlled by other mechanisms in addition to the PTM-regulated nuclear import process. Such kinds of multiple regulation mechanisms are common in biological systems to make the regulated activity tightly controlled, which is also e.g. reported in experimental studies [5]. The first two hypothesized regulation mechanisms regulate biological activities in different stages of gene expression while the third hypothesized regulation mechanism regulates biological activities during cell proliferation. The diversity of the biological activities regulated by PTM-based nuclear import regulations indicates it is widely involved in various biological activities within the cell nucleus.

4.4 Conclusions

In this study, we proposed a computational method, NIpredict, for predicting nuclear import activity and discovery of PTM-based nuclear import regulations. This approach is based on characterizing the interaction between NLS and the import receptor in terms of MIEC and learning the relationship between the characterized interaction energies and the corresponding nuclear import activity by Support Vector regression. The accuracy of NIpredict is demonstrated by its high performance in leave-one-out cross-validation on the training datasets and accurate prediction in the real case. NIpredict was then used to systematically scan the Yeast and Human genome for identifying potential PTM-based nuclear import regulations. Based on NIpredict predictions and known biological roles of the PTMs (or PTM-mediating

enzymes), we identified the potential regulation mechanisms of three biological activities through the identified PTM-based nuclear import regulation. It should be noted that the scope of analysis in this study was limited by the NLS mutation templates due to limited experimental dataset. This approach can be applied to identify more comprehensive list of PTM-based regulations of protein sub-cellular localization given more experimental datasets. A web server for predicting nuclear import activity given the NLS sequence is available at http://mleg.cse.sc.edu/NIpredict.

Chapter 5

Conclusions

My dissertation is composed of three parts which addressed issues in the corresponding topics: prediction of protein localization, prediction of protein sorting signal, and identification of protein localization regulation. They were organized into Chapter 2, Chapter 3 and Chapter 4 respectively. In the next section I will give a brief summary of my dissertation.

5.1 Summary

Protein localization has been recognized as useful information for protein function annotation. Despite many computational methods have been proposed for protein localization prediction, their prediction accuracies are far from being sufficient for genome wide protein localization prediction. Ensemble methods have been proposed as solutions for achieving higher prediction accuracies by combining strength of different protein localization predictors. However, no previous works addressed the issue of intensive computation for applying the ensemble solutions. In the first part of the dissertation, a framework of designing minimalist ensemble algorithms for practical genome-wide protein subcellular localization prediction is proposed, which can significantly reduce the number of individual predictors in a given ensemble algorithm while maintaining comparable performance. In particular, we analyzed the predictions of 9 existing protein localization predictors and addressed issues of algorithm redundancy, consensus mistakes, and algorithm complementarity in designing ensemble

algorithms.

Sorting signals are direct evidences for protein localization. Prediction of sorting signals can thus help elucidate the functions of proteins. In the second part of the dissertation we investigated nuclear localization signals and proposed SeqNLS, a novel computational method for NLS prediction. SeqNLS outperformed other state-of-the-art NLS prediction algorithms for two main reasons: the algorithm can extensively identify potential NLSs by mined NLS sequence patterns through applying frequent pattern mining techniques; SeqNLS incorporates the linear motif attributes of NLS which can effectively remove false positive predictions.

Nuclear import of proteins can be regulated through modulating their NLSs by PTM. In the third part of the dissertation we proposed NIpredict to predict nuclear import activity based on characterized NLS-import receptor interaction. Our experiments showed that nuclear import activity change due to NLS change could be accurately predicted by the NIpredict algorithm. Based on NIpredict, we developed a systematic framework to identify potential PTM-based nuclear import regulations for human and yeast nuclear proteins. Application of this approach has uncovered the potential nuclear import regulation mechanisms by phosphorylation and/or acetylation of three nuclear proteins including SF1, histone H1, and ORC6.

5.2 Main Conclusions

Despite the fact that existing protein localization prediction methods use different algorithms to predict protein localization, many of them use similiar features and thus tend to make wrong predictions on the same proteins. Incorporation of prediction algorithms using distinct features rather than high-performance prediction algorithms contributes to ensemble predictions.

The importance of incorporating distinct features for designing an prediction algorithm is also addressed in Chapter 3. While most other NLS prediction algorithms

simply utilize sequence features, an important factor that SeqNLS outperforms other NLS prediction algorithms is that SeqNLS incorporates the linear motif attributes of NLS in addition to sequence features.

Protein interactions are essential events for controlling cellular processes including protein localization. In Chapter 4, to the best of my knowledge, it is the first time that protein localization is investigated through modeling protein-protein interactions. Despite the limitation of experimental dataset, results of serveral uncovered potential regulation mechanisms have demonstrated its research potential and await further experimental verification.

Bibliography

[1] W. K. Huh, J. V. Falvo, L. C. Gerke, A. S. Carroll, R. W. Howson, J. S. Weissman, and E. K. O'Shea. Global analysis of protein localization in budding yeast. Nature, 425(6959):686–91, 2003.

[2] K. Laurila and M. Vihinen. Prediction of disease-related mutations affecting protein localization. BMC Genomics, 10:122, 2009.

[3] S. P. Wang, T. Q. Guo, X. Y. Guo, J. T. Huang, and C. D. Lu. Structural analysis of fibroin heavy chain signal peptide of silkworm bombyx mori. Acta Biochim Biophys Sin (Shanghai), 38(7):507–13, 2006.

[4] X. Guo, Y. Zhang, X. Zhang, S. Wang, and C. Lu. Recognition of signal peptide by protein translocation machinery in middle silk gland of silkworm bombyx mori. Acta Biochim Biophys Sin (Shanghai), 40(1):38–46, 2008.

[5] A. Kaffman and E. K. O'Shea. Regulation of nuclear localization: a key to a door. Annual review of cell and developmental biology, 15:291–339, 1999.

[6] I. K. Poon and D. A. Jans. Regulation of nuclear transport: central role in development and transformation? Traffic, 6(3):173–86, 2005.

[7] M. N. Chahine and G. N. Pierce. Therapeutic targeting of nuclear protein import in pathological cell conditions. Pharmacological reviews, 61(3):358–72, 2009.

[8] L. M. McLane and A. H. Corbett. Nuclear localization signals and human disease. IUBMB life, 61(7):697–706, 2009.

[9] M. C. Hung and W. Link. Protein localization in disease and therapy. Journal of cell science, 124(Pt 20):3381–92, 2011.

[10] T. R. Kau, J. C. Way, and P. A. Silver. Nuclear transport and cancer: from mechanism to intervention. Nature reviews. Cancer, 4(2):106–17, 2004.

[11] K. Imai and K. Nakai. Prediction of subcellular locations of proteins: where to proceed? Proteomics, 10(22):3970–83, 2010.

[12] J. Sprenger, J. L. Fink, and R. D. Teasdale. Evaluation and comparison of mammalian subcellular localization prediction methods. Bmc Bioinformatics, 7 Suppl 5:S3, 2006.

[13] J. Liu, S. Kang, C. Tang, L. B. Ellis, and T. Li. Meta-prediction of protein subcellular localization with reduced voting. Nucleic Acids Res, 35(15):e96, 2007.

[14] K. Laurila and M. Vihinen. Prolocalizer: integrated web service for protein subcellular localization prediction. Amino Acids, 40(3):975–80, 2011.

[15] S. Park, J. S. Yang, S. K. Jang, and S. Kim. Construction of functional interaction networks through consensus localization predictions of the human proteome. J Proteome Res, 8(7):3367–76, 2009.

[16] J. Assfalg, J. Gong, H. P. Kriegel, A. Pryakhin, T. Wei, and A. Zimek. Supervised ensembles of prediction methods for subcellular localization. J Bioinform Comput Biol, 7(2):269–85, 2009.

[17] Y. Q. Shen and G. Burger. 'unite and conquer': enhanced prediction of protein subcellular localization by integrating multiple specialized tools. Bmc Bioinformatics, 8:420, 2007.

[18] K. T. Lythgow, G. Hudson, P. Andras, and P. F. Chinnery. A critical analysis of the combined usage of protein localization prediction methods: Increasing the number of independent data sets can reduce the accuracy of predicted mitochondrial localization. Mitochondrion, 11(3):444–9, 2011.

[19] K. Nakai and P. Horton. Psort: a program for detecting sorting signals in proteins and predicting their subcellular localization. Trends Biochem Sci, 24(1):34–6, 1999.

[20] M. Cokol, R. Nair, and B. Rost. Finding nuclear localization signals. EMBO Rep, 1(5):411–5, 2000.

[21] A. N. Nguyen Ba, A. Pogoutse, N. Provart, and A. M. Moses. Nlstradamus: a simple hidden markov model for nuclear localization signal prediction. Bmc Bioinformatics, 10:202, 2009.

[22] S. Kosugi, M. Hasebe, M. Tomita, and H. Yanagawa. Systematic identification of cell cycle-dependent yeast nucleocytoplasmic shuttling proteins by prediction of composite motifs. Proc Natl Acad Sci U S A, 106(25):10171–6, 2009.

[23] A. M. Mehdi, M. S. Sehgal, B. Kobe, T. L. Bailey, and M. Boden. A probabilistic model of nuclear import of proteins. Bioinformatics, 27(9):1239–46, 2011.

[24] G. R. Hicks and N. V. Raikhel. Protein import into the nucleus: An integrated view. Annual Review of Cell and Developmental Biology, 11:155–188, 1995.

[25] A. Lange, R. E. Mills, C. J. Lange, M. Stewart, S. E. Devine, and A. H. Corbett. Classical nuclear localization signals: Definition, function, and interaction with importin alpha. Journal of Biological Chemistry, 282(8):5101–5105, 2007.

[26] M. T. Harreman, T. M. Kline, H. G. Milford, M. B. Harben, A. E. Hodel, and A. H. Corbett. Regulation of nuclear import by phosphorylation adjacent to nuclear localization signals. The Journal of biological chemistry, 279(20):20613–21, 2004.

[27] J. D. Nardozzi, K. Lott, and G. Cingolani. Phosphorylation meets nuclear import: a review. Cell communication and signaling : CCS, 8:32, 2010.

[28] D. A. Jans, C. Y. Xiao, and M. H. Lam. Nuclear targeting signal recognition: a key control point in nuclear transport? Bioessays, 22(6):532–44, 2000.

[29] Y. Liu, L. Peng, E. Seto, S. Huang, and Y. Qiu. Modulation of histone deacetylase 6 (hdac6) nuclear import and tubulin deacetylase activity through acetylation. The Journal of biological chemistry, 287(34):29168–74, 2012.

[30] D. L. Madison, P. Yaciuk, R. P. Kwok, and J. R. Lundblad. Acetylation of the adenovirus-transforming protein e1a determines nuclear localization by disrupting association with importin-alpha. The Journal of biological chemistry, 277(41):38755–63, 2002.

[31] M. G. di Bari, L. Ciuffini, M. Mingardi, R. Testi, S. Soddu, and D. Barila. c-abl acetylation by histone acetyltransferases regulates its nuclear-cytoplasmic localization. EMBO reports, 7(7):727–33, 2006.

[32] T. Dietschy, I. Shevelev, J. Pena-Diaz, D. Huhn, S. Kuenzle, R. Mak, M. F. Miah, D. Hess, M. Fey, M. O. Hottiger, P. Janscak, and I. Stagljar. p300-mediated acetylation of the rothmund-thomson-syndrome gene product recql4 regulates its subcellular localization. Journal of cell science, 122(Pt 8):1258–67, 2009.

[33] T. Shimazu, S. Horinouchi, and M. Yoshida. Multiple histone deacetylases and the creb-binding protein regulate pre-mrna 3'-end processing. The Journal of biological chemistry, 282(7):4470–8, 2007.

[34] T. Li, B. A. Diner, J. Chen, and I. M. Cristea. Acetylation modulates cellular distribution and dna sensing ability of interferon-inducible protein ifi16. Proceedings of the National Academy of Sciences of the United States of America, 109(26):10558–63, 2012.

[35] L. Thevenet, C. Mejean, B. Moniot, N. Bonneaud, N. Galeotti, G. Aldrian-Herrada, F. Poulat, P. Berta, M. Benkirane, and B. Boizet-Bonhoure. Regulation of human sry subcellular distribution by its acetylation/deacetylation. The EMBO journal, 23(16):3336–45, 2004.

[36] C. Spilianakis, J. Papamatheakis, and A. Kretsovali. Acetylation by pcaf enhances ciita nuclear accumulation and transactivation of major histocompatibility complex class ii genes. Molecular and cellular biology, 20(22):8489–98, 2000.

[37] B. L. Timney, J. Tetenbaum-Novatt, D. S. Agate, R. Williams, W. Zhang, B. T. Chait, and M. P. Rout. Simple kinetic relationships and nonspecific competition govern nuclear import rates in vivo. The Journal of cell biology, 175(4):579–93, 2006.

[38] A. E. Hodel, M. T. Harreman, K. F. Pulliam, M. E. Harben, J. S. Holmes, M. R. Hodel, K. M. Berland, and A. H. Corbett. Nuclear localization signal receptor affinity correlates with in vivo localization in saccharomyces cerevisiae. The Journal of biological chemistry, 281(33):23545–56, 2006.

[39] S. N. Yang, A. A. Takeda, M. R. Fontes, J. M. Harris, D. A. Jans, and B. Kobe. Probing the specificity of binding to the major nuclear localization sequence-binding site of importin-alpha using oriented peptide library screening. The Journal of biological chemistry, 285(26):19935–46, 2010.

[40] S. Ghosh, A. P. Vassilev, J. Zhang, Y. Zhao, and M. L. DePamphilis. Assembly of the human origin recognition complex occurs through independent nuclear localization of its components. The Journal of biological chemistry, 286(27):23831–41, 2011.

[41] A. Asada, T. Saito, and S. Hisanaga. Phosphorylation of p35 and p39 by cdk5 determines the subcellular location of the holokinase in a phosphorylation-site-specific manner. Journal of cell science, 125(Pt 14):3421–9, 2012.

[42] J. Chung, P. Khadka, and I. K. Chung. Nuclear import of htert requires a bipartite nuclear localization signal and akt-mediated phosphorylation. Journal of cell science, 125(Pt 11):2684–97, 2012.

[43] A. C. Teng, N. A. Al-Montashiri, B. L. Cheng, P. Lou, P. Ozmizrak, H. H. Chen, and A. F. Stewart. Identification of a phosphorylation-dependent nuclear localization motif in interferon regulatory factor 2 binding protein 2. PLoS One, 6(8):e24100, 2011.

[44] P. Fernandez-Garcia, R. Pelaez, P. Herrero, and F. Moreno. Phosphorylation of yeast hexokinase 2 regulates its nucleocytoplasmic shuttling. The Journal of biological chemistry, 287(50):42151–64, 2012.

[45] A. J. Fulcher, D. M. Roth, S. Fatima, G. Alvisi, and D. A. Jans. The brca-1 binding protein brap2 is a novel, negative regulator of nuclear import of viral proteins, dependent on phosphorylation flanking the nuclear localization signal. FASEB journal : official publication of the Federation of American Societies for Experimental Biology, 24(5):1454–66, 2010.

[46] G. Yu, T. Yan, Y. Feng, X. Liu, Y. Xia, H. Luo, J. Z. Wang, and X. Wang. Ser9 phosphorylation causes cytoplasmic detention of i2pp2a/set in alzheimer disease. Neurobiology of aging, 34(7):1748–58, 2013.

[47] T. Hou, W. Zhang, D. A. Case, and W. Wang. Characterization of domain-peptide interaction interface: a case study on the amphiphysin-1 sh3 domain. Journal of Molecular Biology, 376(4):1201–14, 2008.

[48] T. Hou, Z. Xu, W. Zhang, W. A. McLaughlin, D. A. Case, Y. Xu, and W. Wang. Characterization of domain-peptide interaction interface: a generic structure-based model to decipher the binding specificity of sh3 domains. Mol Cell Proteomics, 8(4):639–49, 2009.

[49] T. Hou, W. Zhang, J. Wang, and W. Wang. Predicting drug resistance of the hiv-1 protease using molecular interaction energy components. Proteins, 74(4):837–46, 2009.

[50] N. Li, T. Hou, B. Ding, and W. Wang. Characterization of pdz domain-peptide interaction interface based on energetic patterns. Proteins, 79(11):3208–20, 2011.

[51] J. R. Lin, A. M. Mondal, R. Liu, and J. Hu. Minimalist ensemble algorithms for genome-wide protein localization prediction. Bmc Bioinformatics, 13:157, 2012.

[52] J. R. Lin and J. Hu. Seqnls: Nuclear localization signal prediction based on frequent pattern mining and linear motif scoring. PLoS One, 8(10):e76864, 2013.

[53] J. Assfalg, J. Gong, H. P. Kriegel, A. Pryakhin, T. D. Wei, and A. Zimek. Investigating a correlation between subcellular localization and fold of proteins. Journal of Universal Computer Science, 16(5):604–621, 2010.

[54] S. Briesemeister, J. Rahnenfuhrer, and O. Kohlbacher. Going from where to why–interpretable prediction of protein subcellular localization. Bioinformatics, 26(9):1232–8, 2010.

[55] T. Blum, S. Briesemeister, and O. Kohlbacher. Multiloc2: integrating phylogeny and gene ontology terms improves subcellular protein localization prediction. Bmc Bioinformatics, 10:274, 2009.

[56] H. N. Lin, C. T. Chen, T. Y. Sung, S. Y. Ho, and W. L. Hsu. Protein subcellular localization prediction of eukaryotes using a knowledge-based approach. Bmc Bioinformatics, 10, 2009.

[57] B. Niu, Y. H. Jin, K. Y. Feng, W. C. Lu, Y. D. Cai, and G. Z. Li. Using adaboost for the prediction of subcellular location of prokaryotic and eukaryotic proteins. Molecular Diversity, 12(1):41–45, 2008.

[58] P. Horton, K. J. Park, T. Obayashi, N. Fujita, H. Harada, C. J. Adams-Collier, and K. Nakai. Wolf psort: protein localization predictor. Nucleic Acids Research, 35:W585–W587, 2007.

[59] A. Pierleoni, P. L. Martelli, P. Fariselli, and R. Casadio. Bacello: a balanced subcellular localization predictor. Bioinformatics, 22(14):E408–E416, 2006.

[60] C. S. Yu, Y. C. Chen, C. H. Lu, and J. K. Hwang. Prediction of protein subcellular localization. Proteins-Structure Function and Bioinformatics, 64(3):643–651, 2006.

[61] S. J. Hua and Z. R. Sun. Support vector machine approach for protein subcellular localization prediction. Bioinformatics, 17(8):721–728, 2001.

[62] Ananda MM and Jianjun Hu. Netloc: Network based protein localization prediction using protein-protein interaction and co-expression networks. In BIBM, pages 142–148, 2010.

[63] H. Hishigaki, K. Nakai, T. Ono, A. Tanigami, and T. Takagi. Assessment of prediction accuracy of protein function from protein–protein interaction data. Yeast, 18(6):523–31, 2001.

[64] K. Lee, H. Y. Chuang, A. Beyer, M. K. Sung, W. K. Huh, B. Lee, and T. Ideker. Protein networks markedly improve prediction of subcellular localization in multiple eukaryotic species. Nucleic Acids Res, 36(20):e136, 2008.

[65] C. J. Shin, S. Wong, M. J. Davis, and M. A. Ragan. Protein-protein interaction as a predictor of subcellular location. Bmc Systems Biology, 3, 2009.

[66] C. Stark, B. J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers. Biogrid: a general repository for interaction datasets. Nucleic Acids Research, 34:D535–D539, 2006.

[67] Bongard J Lu XW Z, Zhu X. Ensemble pruning via individual contribution ordering. In Proc of KDD, pages 871–880, 2010.

[68] Hall MA. Correlation-based feature subset selection for machine learning. Dissertation, 1999.

[69] J. Sprenger, J. Lynn Fink, S. Karunaratne, K. Hanson, N. A. Hamilton, and R. D. Teasdale. Locate: a mammalian protein subcellular localization database. Nucleic Acids Res, 36(Database issue):D230–3, 2008.

[70] M. Marfori, A. Mynott, J. J. Ellis, A. M. Mehdi, N. F. Saunders, P. M. Curmi, J. K. Forwood, M. Boden, and B. Kobe. Molecular basis for specificity of nuclear import and prediction of nuclear localization. Biochim Biophys Acta, 1813(9):1562–77, 2011.

[71] P. Puntervoll, R. Linding, C. Gemund, S. Chabanis-Davidson, M. Mattingsdal, S. Cameron, D. M. A. Martin, G. Ausiello, B. Brannetti, A. Costantini, F. Ferre, V. Maselli, A. Via, G. Cesareni, F. Diella, G. Superti-Furga, L. Wyrwicz, C. Ramu, C. McGuigan, R. Gudavalli, I. Letunic, P. Bork, L. Rychlewski, B. Kuster, M. Helmer-Citterich, W. N. Hunter, R. Aasland, and T. J. Gibson.

Elm server: a new resource for investigating short functional sites in modular eukaryotic proteins. Nucleic Acids Research, 31(13):3625–3630, 2003.

[72] F. Diella, N. Haslam, C. Chica, A. Budd, S. Michael, N. P. Brown, G. Trave, and T. J. Gibson. Understanding eukaryotic linear motifs and their role in cell signaling and regulation. Frontiers in Bioscience, 13:6580–6603, 2008.

[73] V. Neduva and R. B. Russell. Linear motifs: Evolutionary interaction switches. Febs Letters, 579(15):3342–3345, 2005.

[74] A. Bairoch. Prosite - a dictionary of sites and patterns in proteins. Nucleic Acids Research, 20:2013–2018, 1992.

[75] J. C. Obenauer, L. C. Cantley, and M. B. Yaffe. Scansite 2.0: proteome-wide prediction of cell signaling interactions using short sequence motifs. Nucleic Acids Research, 31(13):3635–3641, 2003.

[76] S. Balla, V. Thapar, S. Verma, T. Luong, T. Faghri, C. H. Huang, S. Rajasekaran, J. J. del Campo, J. H. Shinn, W. A. Mohler, M. W. Maciejewski, M. R. Gryk, B. Piccirillo, S. R. Schiller, and M. R. Schiller. Minimotif miner: a tool for investigating protein function. Nature Methods, 3(3):175–177, 2006.

[77] R. Gutman, C. Berezin, R. Wollman, Y. Rosenberg, and N. Ben-Tal. Quasimotifinder: protein annotation by searching for evolutionarily conserved motif-like patterns. Nucleic Acids Research, 33:W255–W261, 2005.

[78] D. Plewczynski, A. Tkacz, L. S. Wyrwicz, and L. Rychlewski. Automotif server: prediction of single residue post-translational modifications in proteins. Bioinformatics, 21(10):2525–2527, 2005.

[79] C. Ramu. Sirw: a web server for the simple indexing and retrieval system that combines sequence motif searches with keyword searches. Nucleic Acids Research, 31(13):3771–3774, 2003.

[80] N. E. Davey, N. J. Haslam, D. C. Shields, and R. J. Edwards. Slimsearch 2.0: biological context for short linear motifs in proteins. Nucleic Acids Research, 39:W56–W60, 2011.

[81] V. Neduva and R. B. Russell. Dilimot: discovery of linear motifs in proteins. Nucleic Acids Research, 34:W350–W355, 2006.

[82] R. J. Edwards, N. E. Davey, and D. C. Shields. Slimfinder: A probabilistic method for identifying over-represented, convergently evolved, short linear motifs in proteins. Plos One, 2(10), 2007.

[83] S. H. Tan, W. Hugo, W. K. Sung, and S. K. Ng. A correlated motif approach for finding short linear motifs from protein interaction networks. Bmc Bioinformatics, 7, 2006.

[84] N. J. Haslam and D. C. Shields. Profile-based short linear protein motif discovery. Bmc Bioinformatics, 13, 2012.

[85] N. E. Davey, K. Van Roey, R. J. Weatheritt, G. Toedt, B. Uyar, B. Altenberg, A. Budd, F. Diella, H. Dinkel, and T. J. Gibson. Attributes of short linear motifs. Molecular Biosystems, 8(1):268–281, 2012.

[86] R. Nair, P. Carter, and B. Rost. Nlsdb: database of nuclear localization signals. Nucleic Acids Research, 31(1):397–399, 2003.

[87] S. Kosugi, M. Hasebe, N. Matsumura, H. Takashima, E. Miyamoto-Sato, M. Tomita, and H. Yanagawa. Six classes of nuclear localization signals specific to different binding grooves of importin alpha. Journal of Biological Chemistry, 284(1):478–485, 2009.

[88] B. J. Lee, A. E. Cansizoglu, K. E. Suel, T. H. Louis, Z. Zhang, and Y. M. Chook. Rules for nuclear localization sequence recognition by karyopherin beta 2. Cell, 126(3):543–58, 2006.

[89] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. Introduction to data mining. Pearson Addison Wesley, Boston, 1st edition, 2006.

[90] D. A. Jans, C. Y. Xiao, and M. H. C. Lam. Nuclear targeting signal recognition: a key control point in nuclear transport? Bioessays, 22(6):532–544, 2000.

[91] S. Hahn, P. Maurer, S. Caesar, and G. Schlenstedt. Classical nls proteins from saccharomyces cerevisiae. Journal of Molecular Biology, 379(4):678–694, 2008.

[92] A. Lange, L. M. McLane, R. E. Mills, S. E. Devine, and A. H. Corbett. Expanding the definition of the classical bipartite nuclear localization signal. Traffic, 11(3):311–323, 2010.

[93] J. Robbins, S. M. Dilworth, R. A. Laskey, and C. Dingwall. Two interdependent basic domains in nucleoplasmin nuclear targeting sequence: identification of a class of bipartite nuclear targeting sequence. Cell, 64(3):615–23, 1991.

[94] M. R. Fontes, T. Teh, D. Jans, R. I. Brinkworth, and B. Kobe. Structural basis for the specificity of bipartite nuclear localization sequence binding by importin-alpha. Journal of Biological Chemistry, 278(30):27981–7, 2003.

[95] M. Fuxreiter, P. Tompa, and I. Simon. Local structural disorder imparts plasticity on linear motifs. Bioinformatics, 23(8):950–6, 2007.

[96] T. Ishida and K. Kinoshita. Prdos: prediction of disordered protein regions from amino acid sequence. Nucleic Acids Research, 35(Web Server issue):W460–4, 2007.

[97] B. Monastyrskyy, K. Fidelis, J. Moult, A. Tramontano, and A. Kryshtafovych. Evaluation of disorder predictions in casp9. Proteins, 79 Suppl 10:107–18, 2011.

[98] A. Via, C. M. Gould, C. Gemund, T. J. Gibson, and M. Helmer-Citterich. A structure filter for the eukaryotic linear motif resource. Bmc Bioinformatics, 10:351, 2009.

[99] B. Petersen, T. N. Petersen, P. Andersen, M. Nielsen, and C. Lundegaard. A generic method for assignment of reliability scores applied to solvent accessibility predictions. BMC Struct Biol, 9:51, 2009.

[100] R. E. Fan, P. H. Chen, and C. J. Lin. Working set selection using second order information for training support vector machines. Journal of Machine Learning Research, 6:1889–1918, 2005.

[101] N. E. Davey, D. C. Shields, and R. J. Edwards. Masking residues using context-specific evolutionary conservation significantly improves short linear motif discovery. Bioinformatics, 25(4):443–50, 2009.

[102] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. H. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. Nucleic Acids Research, 25(17):3389–3402, 1997.

[103] M. Tompa, N. Li, T. L. Bailey, G. M. Church, B. De Moor, E. Eskin, A. V. Favorov, M. C. Frith, Y. T. Fu, W. J. Kent, V. J. Makeev, A. A. Mironov, W. S. Noble, G. Pavesi, G. Pesole, M. Regnier, N. Simonis, S. Sinha, G. Thijs, J. van Helden, M. Vandenbogaert, Z. P. Weng, C. Workman, C. Ye, and Z. Zhu. Assessing computational tools for the discovery of transcription factor binding sites. Nature Biotechnology, 23(1):137–144, 2005.

[104] S. C. Fu, K. Imai, and P. Horton. Prediction of leucine-rich nuclear export signal containing proteins with nessential. Nucleic acids research, 39(16):e111, 2011.

[105] L. I. Davis. The nuclear pore complex. Annual review of biochemistry, 64:865–96, 1995.

[106] A. V. Sorokin, E. R. Kim, and L. P. Ovchinnikov. Nucleocytoplasmic transport of proteins. Biochemistry. Biokhimiia, 72(13):1439–57, 2007.

[107] M. Marfori, T. G. Lonhienne, J. K. Forwood, and B. Kobe. Structural basis of high-affinity nuclear localization signal interactions with importin-alpha. Traffic, 13(4):532–48, 2012.

[108] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The protein data bank. Nucleic acids research, 28(1):235–42, 2000.

[109] Y. Yang, J. Zhan, H. Zhao, and Y. Zhou. A new size-independent score for pairwise protein structure alignment and its application to structure classification and nucleic-acid binding prediction. Proteins, 80(8):2080–8, 2012.

[110] G. G. Krivov, M. V. Shapovalov, and Jr. Dunbrack, R. L. Improved prediction of protein side-chain conformations with scwrl4. Proteins, 77(4):778–95, 2009.

[111] Y. Duan, C. Wu, S. Chowdhury, M. C. Lee, G. Xiong, W. Zhang, R. Yang, P. Cieplak, R. Luo, T. Lee, J. Caldwell, J. Wang, and P. Kollman. A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. Journal of computational chemistry, 24(16):1999–2012, 2003.

[112] N. Guex and M. C. Peitsch. Swiss-model and the swiss-pdbviewer: an environment for comparative protein modeling. Electrophoresis, 18(15):2714–23, 1997.

[113] A. Onufriev, D. Bashford, and D. A. Case. Modification of the generalized born model suitable for macromolecules. Journal of Physical Chemistry B, 104(15):3712–3720, 2000.

[114] A. Onufriev, D. Bashford, and D. A. Case. Exploring protein native states and large-scale conformational changes with a modified generalized born model. Proteins-Structure Function and Bioinformatics, 55(2):383–394, 2004.

[115] J. W. Craft and G. B. Legge. An amber/dyana/molmol phosphorylated amino acid library set and incorporation into nmr structure calculations. Journal of Biomolecular Nmr, 33(1):15–24, 2005.

[116] G. V. Papamokos, G. Tziatzos, D. G. Papageorgiou, S. D. Georgatos, A. S. Politou, and E. Kaxiras. Structural role of rks motifs in chromatin interactions: A molecular dynamics study of hp1 bound to a variably modified histone tail. Biophysical Journal, 102(8):1926–1933, 2012.

[117] M. R. Hodel, A. H. Corbett, and A. E. Hodel. Dissection of a nuclear localization signal. J Biol Chem, 276(2):1317–25, 2001.

[118] G. Alvisi, D. A. Jans, J. Guo, L. A. Pinna, and A. Ripalti. A protein kinase ck2 site flanking the nuclear targeting signal enhances nuclear transport of human cytomegalovirus ppul44. Traffic, 6(11):1002–13, 2005.

[119] G. Alvisi, O. Marin, G. Pari, M. Mancini, S. Avanzi, A. Loregian, D. A. Jans, and A. Ripalti. Multiple phosphorylation sites at the c-terminus regulate nuclear import of hcmv dna polymerase processivity factor ppul44. Virology, 417(2):259–67, 2011.

[120] Z. Dosztanyi, V. Csizmok, P. Tompa, and I. Simon. Iupred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. Bioinformatics, 21(16):3433–4, 2005.

[121] C. T. Lu, K. Y. Huang, M. G. Su, T. Y. Lee, N. A. Bretana, W. C. Chang, Y. J. Chen, and H. D. Huang. Dbptm 3.0: an informative resource for investigating

substrate site specificity and functional association of protein post-translational modifications. Nucleic acids research, 41(Database issue):D295–305, 2013.

[122] X. Wang, S. Bruderer, Z. Rafi, J. Xue, P. J. Milburn, A. Kramer, and P. J. Robinson. Phosphorylation of splicing factor sf1 on ser20 by cgmp-dependent protein kinase regulates spliceosome assembly. The EMBO journal, 18(16):4549–59, 1999.

[123] A. Vaquero, M. Scher, D. Lee, H. Erdjument-Bromage, P. Tempst, and D. Reinberg. Human sirt1 interacts with histone h1 and promotes formation of facultative heterochromatin. Molecular cell, 16(1):93–105, 2004.

[124] V. Q. Nguyen, C. Co, and J. J. Li. Cyclin-dependent kinases prevent dna re-replication through multiple mechanisms. Nature, 411(6841):1068–73, 2001.

[125] B. M. Green, R. J. Morreale, B. Ozaydin, J. L. Derisi, and J. J. Li. Genome-wide mapping of dna synthesis in saccharomyces cerevisiae reveals that mechanisms preventing reinitiation of dna replication are not redundant. Molecular biology of the cell, 17(5):2401–14, 2006.

[126] R. E. Tanny, D. M. MacAlpine, H. G. Blitzblau, and S. P. Bell. Genome-wide analysis of re-replication reveals inhibitory controls that target multiple stages of replication initiation. Molecular biology of the cell, 17(5):2415–23, 2006.

[127] S. Chen and S. P. Bell. Cdk prevents mcm2-7 helicase loading by inhibiting cdt1 interaction with orc6. Genes & development, 25(4):363–72, 2011.

[128] L. J. Holt, B. B. Tuch, J. Villen, A. D. Johnson, S. P. Gygi, and D. O. Morgan. Global analysis of cdk1 substrate phosphorylation sites provides insights into evolution. Science, 325(5948):1682–6, 2009.