学位論文

# ロイシンリッチ核外移行シグナルに関する研究：

# データセットの構築及び新規特徴量解析による予測法の改良

(Better understanding and recognition of leucine-rich nuclear export signals: expanded dataset, novel feature analysis, and the development of an improved prediction method)

傅 思縉

# Abstract

Protein sub-cellular localization is an important feature and has been commonly used to support many functional hypotheses. The leucine-rich nuclear export signal (NES) is an important sub-cellular targeting signal, which is involved in processes such as signal transduction and cell cycle regulation. Although 15 years has passed since its discovery, limited structural information and high sequence diversity have hampered understanding of this signal. A consensus sequence was proposed based on early examples, but later evidence demonstrated its low sensitivity (~37%). To raise the sensitivity, a more general consensus sequence has been widely used at a cost of greatly increased spurious matches. Despite continued interest amongst molecular biologists in the function and regulation of NES-containing proteins, further bioinformatic characterization of this import signal remains at a standstill. Indeed, most of the recently discovered NES sites have been identified by the consensus sequence despite its unsatisfactory trade-off. On the other hand, the NetNES server provides the only computational method currently available. Although these two methods have been widely used to attempt to find the correct NES position within potential NES-containing proteins, their performance has not yet been evaluated on the basic task of discriminating NES-containing proteins from other proteins. To better characterize the NES, we propose a new approach, NESsential, not only capable of finding the correct position of many NES's at the site level, but potential NES-containing proteins at the protein level. We also collected 70 NES-containing proteins recently discovered to update the dataset to approximately two-fold larger than NESbase, the largest previously available dataset.

***Keywords***

Leucine-rich nuclear export signals, CRM1/Exportin 1 mediated export pathway, sequence analysis

# Acknowledgements

Right after I typed the first word of this dissertation, I could not stop recalling those great helps I have received and kept asking myself: What if my kindly and always energetic advisor, Prof. Paul Horton, did not introduce me to this interesting field and patiently instruct me in this issue? What if all the members in this lab did not spend their valuable time discussing with me? How could I concentrate in this study without my families' support? This dissertation would not be what it is without your helps and involvements. Like the old saying says "man proposes and god disposes", I would like to thank god foremost for his arrangement of our encounter, of which I believe the p-value should be extremely low.

# Contents

# List of Figures



# List of Tables

# Chapter 1

# Introduction

Amongst the complicated "route map" of protein sub-cellular localization, the nucleocytoplasmic traffic of proteins occurs through the nuclear pore complexes (NPC), which allow passive diffusion of small proteins (<40kDa) but require active, i.e. energy-dependent transportation, for larger proteins. The active nucleocytoplasmic pathways are mostly mediated by karyopherin proteins and the specific sequence signals of cargo molecules; nuclear localization signals (NLS) and nuclear export signals (NES), for each direction respectively. Compared with classical NLS's, the classical "leucine-rich" NES's are more difficult to identify correctly because the NES consensus sequence often spuriously matches regions forming the hydrophobic core of proteins [1]. The karyopherin Exportin 1/CRM1(chromosomal region maintenance 1) mediates the export of many cellular and viral proteins containing leucine-rich NES's (Figure 1.1). To date, there are over 75 proteins containing this leucine-rich NES verified experimentally. Many of them are related to signal transduction, cell cycle regulation and, moreover, the export of unspliced or partially spliced viral mRNA such as the HIV-1 Rev protein. Recently, this export pathway has also been suggested to be involved in the mechanism inducing the abnormal localization of many tumor suppressors containing leucine-rich NES's, p53 for instance, in various cancer cells [2].

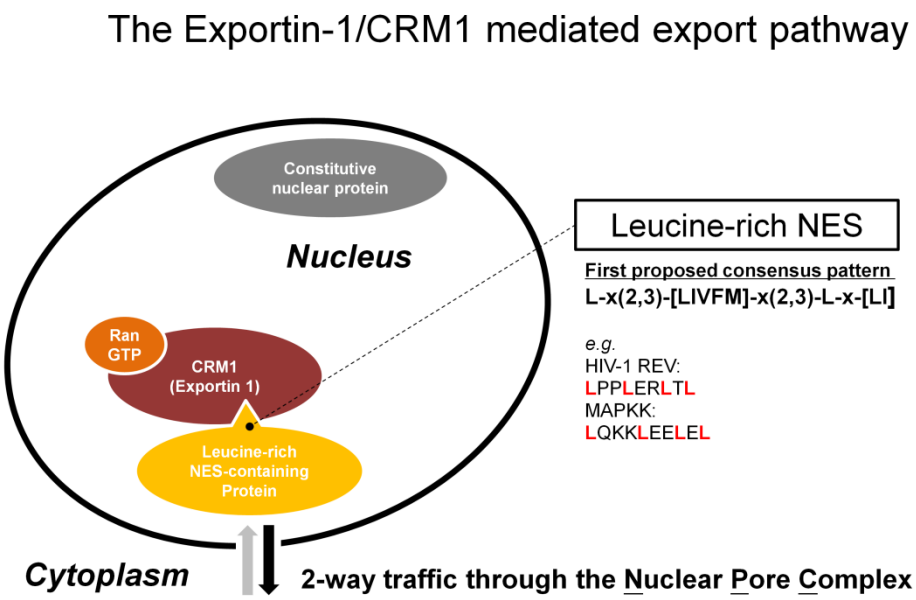



Figure 1.1: The Exportin-1/CRM1 mediated export pathway.

Contrary to its growing importance, we know little about this CRM1-meditated leucine-rich nuclear export signal (NES, hereafter), other than the abundance of hydrophobic residues, mostly leucine, and the specific spacing between them. Limited structural information is one factor which hampers further characterization of the NES. Based on 2nd structure prediction and eight structures (six determined by X-ray crystallography) of NES-containing proteins, previous research had suggested a strong preference of alpha-helical structure and a bias against beta-strands in the N-terminal end of NES's. However, in 2007, the first NES consisting entirely of a beta-sheet was reported in Fibroblast Growth Factor-1 (FGF-1) [3]. Figure 1.2 (Figure 7B from the original paper) shows the diversity of local structure of NES sites.

Unfortunately, no complete structures are available for CRM1 bound to classical NES containing proteins. However, in 2009, the crystal structure of CRM1 in complex with snurpotin 1 (SNUPN), an export substrate previously considered to be exported through an NES-independent interaction with CRM1, was reported [4][5]. This complex structure revealed some details of the binding interface including a minor binding patch near the N-terminus of SNUPN resembling the classical NES. However, this NES mimic may not be sufficient enough to understand the classical NES, because of its much lower binding

affinity. Furthermore, the multipartite recognition and the number of critical hydrophobic residues within this NES mimic are different than what is known about the classical NES's and so far observed only in this CRM1-SNUPN complex.

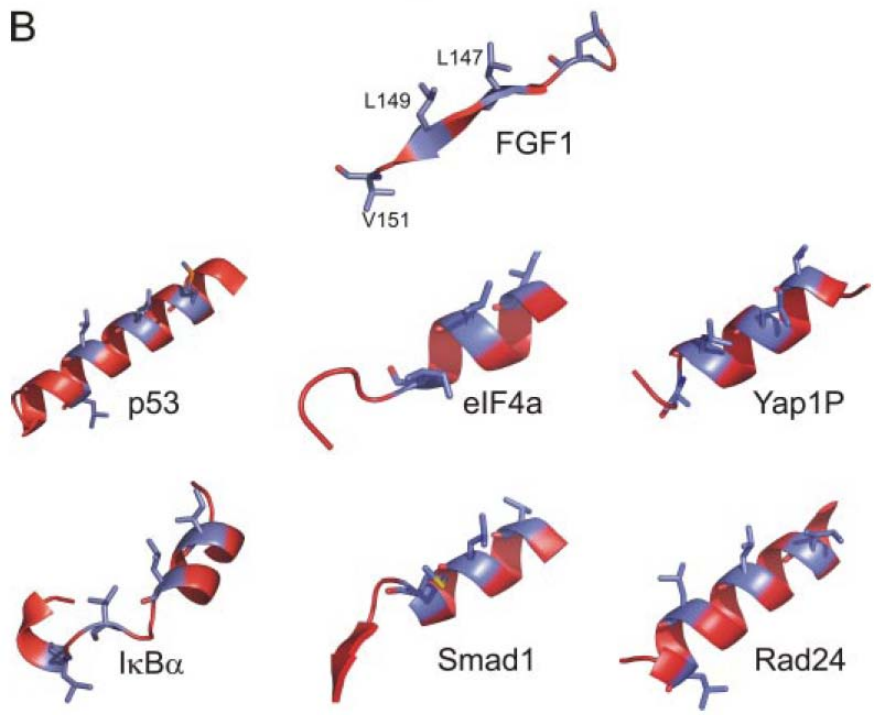



Figure 1.2: The diversity of local structure of NES sites.

The first proposed consensus sequence of this classical NES is L-x-(2,3)-[LIVFM]-x(2,3)-L-x-[LI] where x is any amino acid, defined from analysis of mutant variants of HTLV-1 Rex and HIV-1 Rev [6] following the discovery of NES in the human immunodeficiency virus type 1 (HIV-1) Rev protein [7] and cAMP-dependent protein kinase inhibitor (PKI) [8]. This consensus sequence had been widely used until la Cour et al. indicated that the majority of NES's (63%) in NESbase, a database collecting experimentally verified NES's, deviated from this consensus sequence [9]. Given that the consensus sequence was originally defined by mutant variants of only two proteins, it's not surprising that the incapacity of this consensus sequence has become increasingly evident as more NES-containing proteins are verified. By allowing a more general consensus sequence,

[LIVFM]-x-(2,3)-[LIVFM]-x(2,3)-[LIVFM]-x-[LIVFM], sensitivity can be improved (from 37% to 72%) at a cost of greatly increasing false positives (precision dropped from 52% to 16%) [10]. In practice, this "tolerant" consensus sequence has then been commonly accepted though such a trade-off seems to be unsatisfactory (see Table 1.1).

Table 1.1: The trade-off between sensitivity and precision for each consensus sequence

| | First proposed consensus sequence | More general consensus sequence |
|---|---|---|
| **Sensitivity** | 37% | 72% |
| **Precision** | 52% | 16% |

*Sensitivity: the proportion of NES regions that contain the consensus sequence.
*Precision: the proportion of consensus sequences that are found within NES regions.

Based on NESbase, la Cour et al provided the NetNES web server [10] aiming to solve this condition. The prediction of the NetNES server is performed from primary sequence and implemented by the combination of a hidden Markov model (HMM) and a neural network. Instead of the site-level recognition obtained by consensus sequence match searches, NetNES performs NES prediction at the residue level. Tested on a small set of independent NES-containing proteins, three out of five NES's were correctly located by NetNES. Despite the growing number of experimentally verified NES-containing proteins in recent years, NESbase has stopped updating since 2003. Thus there is an urgent need to collect the NES-containing proteins discovered since then, not only to re-evaluate NetNES, but also to provide a more complete dataset for public use.

Kosugi et al. developed an essay to detect NES's, and proposed an alternative set of consensus sequences [11]. However, they didn't evaluate the trade-off between sensitivity and precision, and in fact their precision is even lower than the more tolerant consensus sequence mentioned above (see Figure 2.4).

Besides finding the correct position of NES's within NES-containing proteins as NetNES and consensus-based methods attempt to do, classification between NES-containing proteins and non-NES-containing proteins is another important issue which hasn't been addressed yet (Figure 1.3). It's more challenging to tackle this issue since some protein-protein interaction domains, such as leucine-rich repeats, also fit the consensus sequence perfectly. Therefore it is important to quantitatively evaluate how effective

previous prediction methods are for this important task.

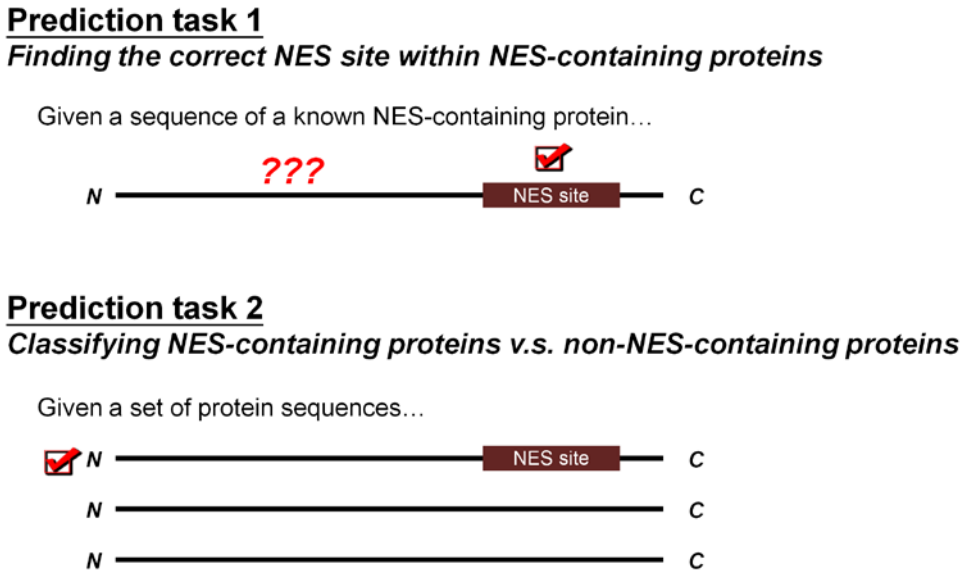



Figure 1.3: Two important prediction tasks for NES's and NES-containing proteins.

In their effort to better characterize NES's, la Cour et al. noted some correlations not included in the consensus sequence representation. In particular, they hypothesized that some protein attributes, such as flexibility, and a minor preference for negative-charged or polar amino acids around the NES are potentially relevant to NES function. Instead of directly using predicted flexibility as a feature of NES's, la Cour et al. built NetNES using primary sequence information alone, perhaps due to the lack of suitable predictors.

Recent research indicates that intrinsic disordered region of proteins are often involved in molecular recognition with both high specificity and low affinity [12][13]. Interestingly, the binding affinities between NES's and CRM1 were found to be generally low and, furthermore, high-affinity artificial NES's impair the efficient release of export complexes from the NPC [14].

In this study, we hypothesized that protein intrinsic disorder may be relevant to NES recognition. We investigated the correlation between protein intrinsic disorder and NES sites and applied our findings to develop a new predictor, NESsential, which aims to not only find the correct position of NES's at the site level, but also potential NES-containing proteins at the protein level.

# Chapter 2

# Material and Methods

## 2.1 *Training data*

We selected 60 NES-containing proteins from NESbase as our training data after the removal of redundant sequences (with sequence identity > 25%), and those lacking of experimental data on CRM1 dependency or sensitivity to leptomycin B (LMB), an effective CRM1 inhibitor, to verify the CRM1-mediated export pathway. The number of training data we use is slightly less than that of NetNES (64 NES-containing proteins from NESbase) due to the stricter identity criteria we applied for NESsential.

## 2.2 *Protein intrinsic disorder prediction*

To investigate the correlation between protein intrinsic disorder and NES function, we applied Poodle-L [15] and DISOPRED [16], two of the best performing tools for disorder prediction in the critical assessment of techniques for protein structure prediction (CASP7), to all protein sequences in the training data. We use both tools to analyze the correlation between intrinsic disorder and NES function, but for prediction we only report the result using POODLE-L, as this choice yielded better NES prediction performance (see Figure S1).

## 2.3 *Training and prediction pipeline of NESsential*

We first applied a pre-filter consensus Φx(2,3)ΦxΦ, where Φ can be substituted by I, V, F or M and x is any amino acid, to each protein sequence in the training set retrieving 946 matches, including 117 NES sites and 829 spurious matches according to the annotation of experimentally verified NES regions. These matches constitute our positive and negative training examples. Subsequently, 22 features, such as predicted disorder, were extracted and applied to train SVM models (implemented by LIBSVM 2.9 [17]) to discriminate between the real NES sites from the spurious matches. LIBSVM package extends SVM and

provides probability estimate besides class label (see supplementary) which is further used in evaluation. Based on this pipeline, we proposed two types of NESsential which differ by a prior classification of matches by disorder prediction: the "flat" NESsential contains one SVM model trained by all matches, while the "split" NESsential employs different SVM models for disordered and ordered matches as shown in Figure 2.3.

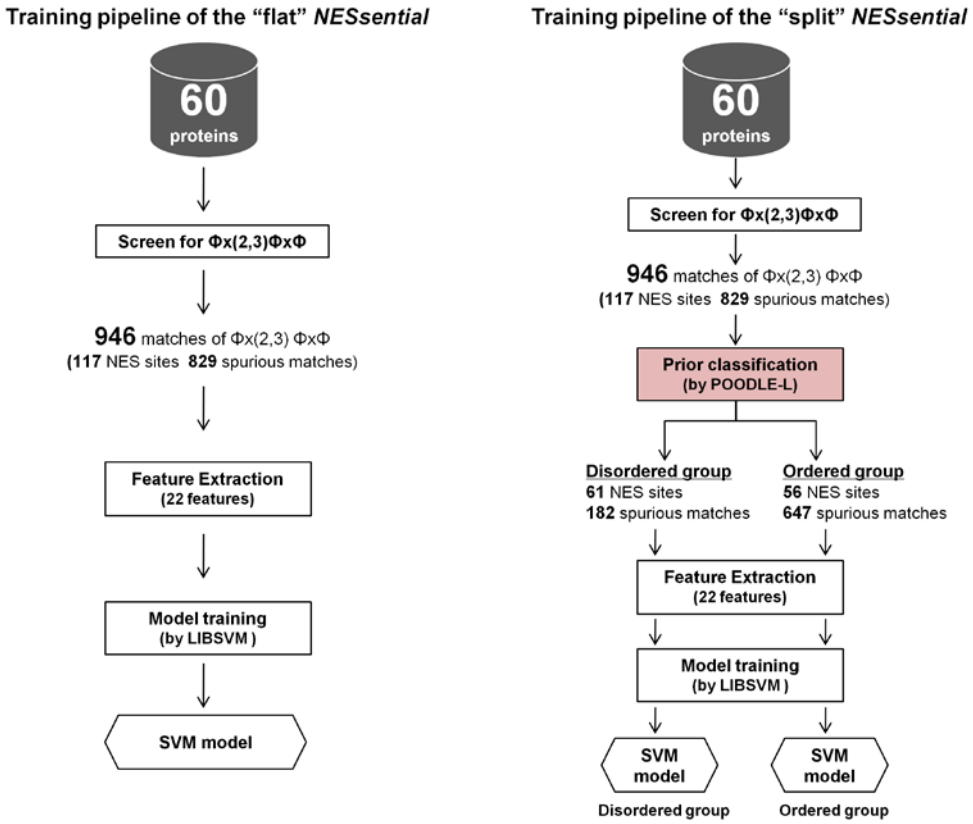



Figure 2.3: The training pipelines of two types of NESsential. In split NESsential, a prior classification is conducted according to the disorder status (by POODL-L) within the sequence matching the pre-filter. If every residue is predicted to be ordered, the pre-filter match will be classified to ordered group; otherwise, disordered group.

## 2.4 *Choice of pre-filters*

In our scheme, it is imperative that a pre-filter have high sensitivity. A low precision is tolerable because the SVM classifier has a chance to eliminate false positives. To increase the transparency of the prediction process, it is also desirable for a pre-filter to be a simple pattern. For these reasons, we applied two general patterns with lengths of 6 and 7 residues, ΦxxΦxΦ and ΦxxxΦxΦ, as a pre-filter. This pre-filter achieve the highest sensitivity amongst all available consensus sequences (Figure 2.4). Moreover, both patterns contain the region bounded by the second and the fourth hydrophobic position of the consensus sequence currently in use. Previous research indicated that the first hydrophobic position in the signal is less conserved [10], which is consistent with some experimental observation indicating the NES activity is more susceptible to mutations of the C-terminal hydrophobic residues within the signal [8][18]. To test the statistical significance, we computed the p-values for these two patterns and obtained p-values of 1.7e-16 (ΦxxΦxΦ) and 5.6e-7 (ΦxxxΦxΦ) respectively. These p-values indicated the probability of finding the two patterns within the verified NES functional regions merely by chance.

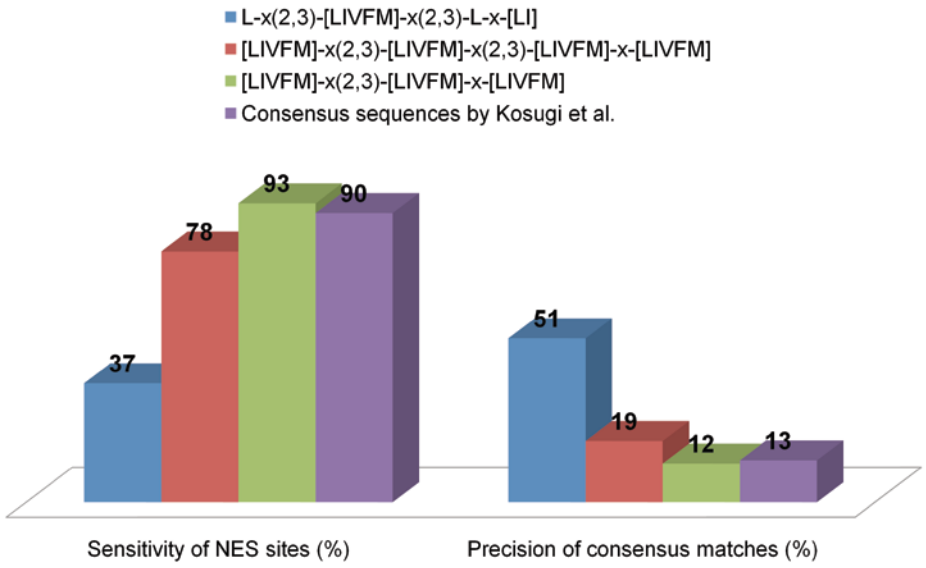



Figure 2.4: The trade-off between different consensus sequences.

# Chapter 3

# Results and Discussion

## 3.1 *Expanding the dataset*

### 3.1.1 Recently discovered NES-containing proteins

To further understanding of NES's and evaluate the performance of existing methods, it is important to use as much experimental data as possible. Starting with the references given by Kosugi et al. [11], we undertook a literature search to collect NES-containing proteins which have been recently discovered and therefore not included in NESbase. In order to allow a fair comparison between our predictor and previous methods, these proteins were used only for evaluation, not training. Sequence identity (25%) was checked by BLASTCLUST to avoid redundancy between training and test data. As a result, we obtained a test set containing 70 proteins and 85 NES's (some proteins contain multiple NES sites). The addition of these newly collected proteins more than doubles the number of CRM1-dependent NES containing proteins organized in a single dataset. This dataset is itself an important resource which should contribute to future NES research (Table S1).

### 3.1.2 Background proteins

The 70 NES-containing proteins described in the previous section can serve as positive examples for protein level prediction. Unfortunately, it is difficult to prepare an ideal negative dataset, as in general it is difficult to rule out the possibility that a nuclear protein may have a yet undiscovered NES or that a non-nuclear protein might have a cryptic NES which could function if the protein were found in the nucleus. Therefore, we selected 541 yeast proteins currently annotated as either "cytosolic-located" (159 proteins) or "nuclear-located" (382 proteins) from the Universal Protein Resource (UniProt) [http://www.uniprot.org/] as background proteins for protein level classification evaluation. A few of these background proteins might contain NES's, but we expect that

most do not. Note that we only use these background proteins for evaluation, never for training.

## 3.2 *Analysis of disorder prediction results*

### 3.2.1 Distribution of predicted disorder scores

Both DISOPRED and POODLE-L return a probability estimate of each residue being disordered. Figure 3.2.1 shows the different distribution of average predicted disorder scores between NES sites and spurious matches. The average scores from both disorder predictors are generally higher within or flanking the NES sites than in spurious matches. Although the scores predicted by POODLE-L show a larger difference between real NES and spurious sites, it should be mentioned that different cutoff values were used for each predictor. POODLE-L uses a simple score threshold of 0.5, but DISOPRED considers both the estimated false positive rate and the predicted disorder score. Thus the predicted disorder scores from these two predictors are not directly comparable.

### 3.2.2 Distribution of predicted disorder scores for NES sites by POODLE-L

Focusing on the predicted disorder scores made by POODLE-L, we found that the average predicted scores within or flanking the NES sites were generally close to the cutoff value of 0.5. The histograms in Figure 3.2.2 show the distribution of predicted disorder scores at position 0, the first residue of the ΦxxΦxΦ and ΦxxxΦxΦ matches respectively. The three peaks found in the distribution for NES sites, suggested that the predicted disorder scores tend to be either lower than 0.2 or higher than 0.6, implying that this position can be either very ordered or very disordered. Similar results were observed at other positions within the pre-filter region. We did not apply this analysis to the predicted disorder scores made by DISOPRED, due to the two degrees of freedom, i.e. probability estimate and false positive rate mentioned in the previous section.

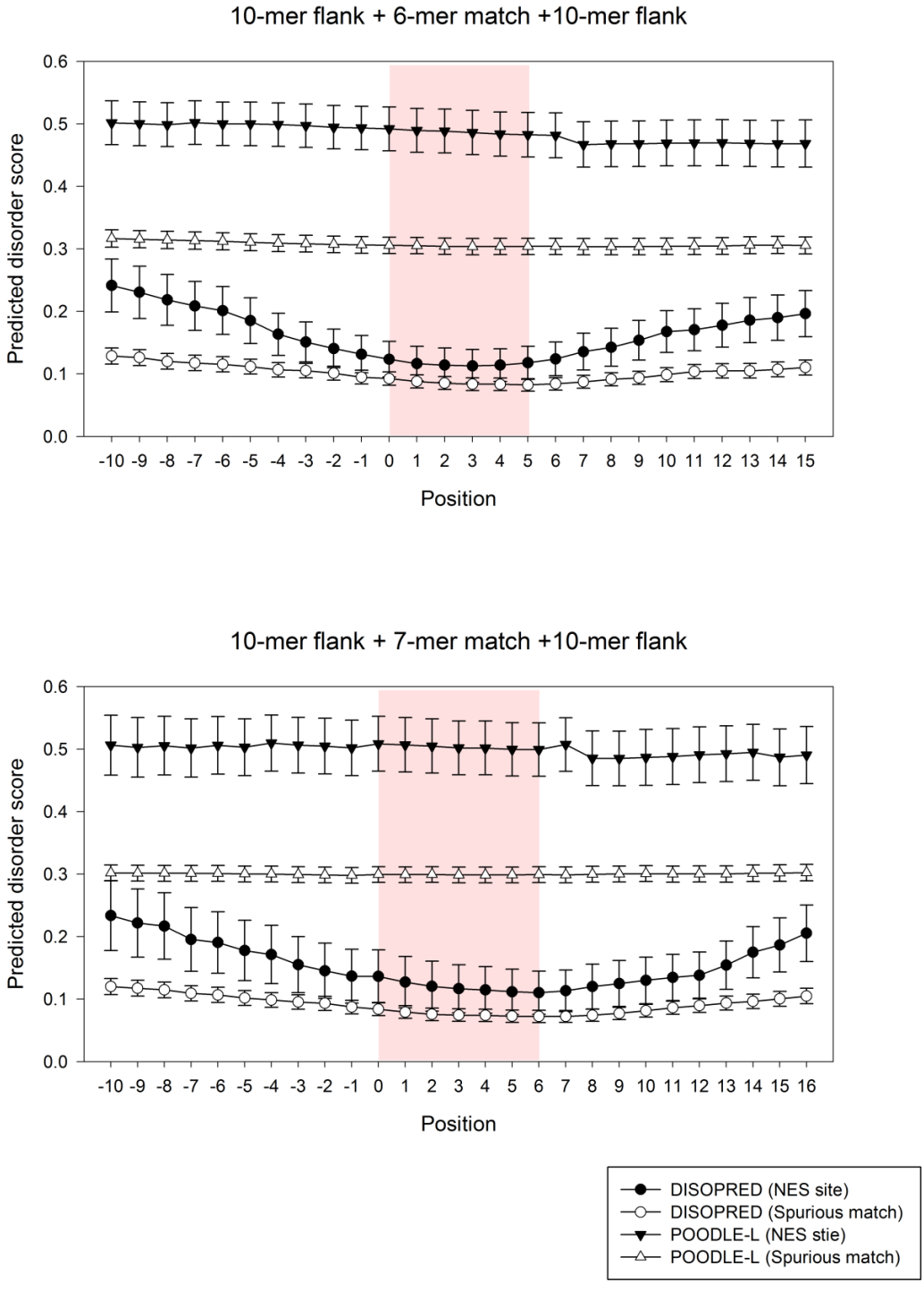
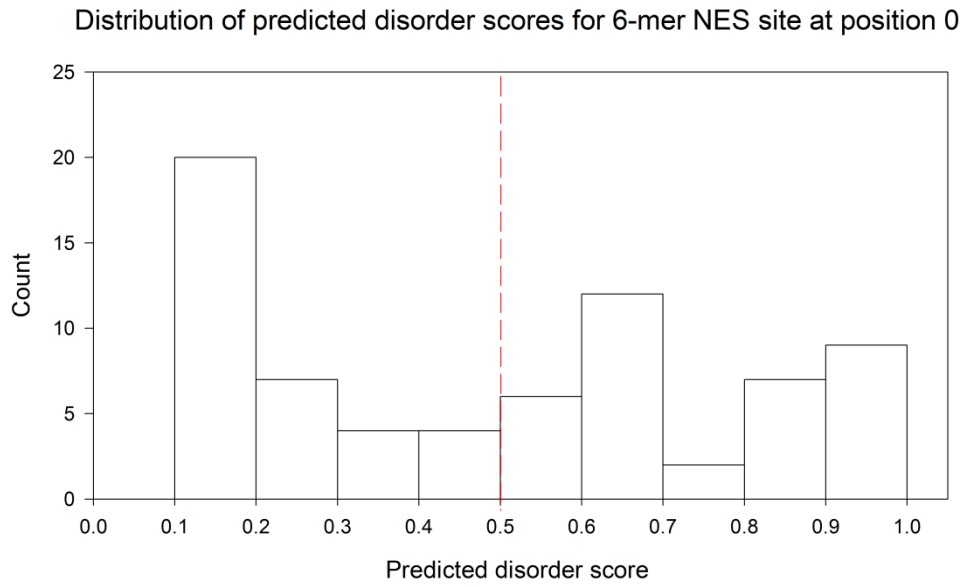



Figure 3.2.1: Analysis of predicted disorder scores by POODLE-L and DISOPRED. Distribution of the mean score and its standard error are shown at each position, where position 0 represents for the first residue of match of 6-mer (ΦxxΦxΦ) or 7-mer (ΦxxxΦxΦ) pre-filter. The regions corresponding to ΦxxΦxΦ and ΦxxxΦxΦ are highlighted in pink. (where Φ denotes [LIVFM] and x denotes any amino acid).



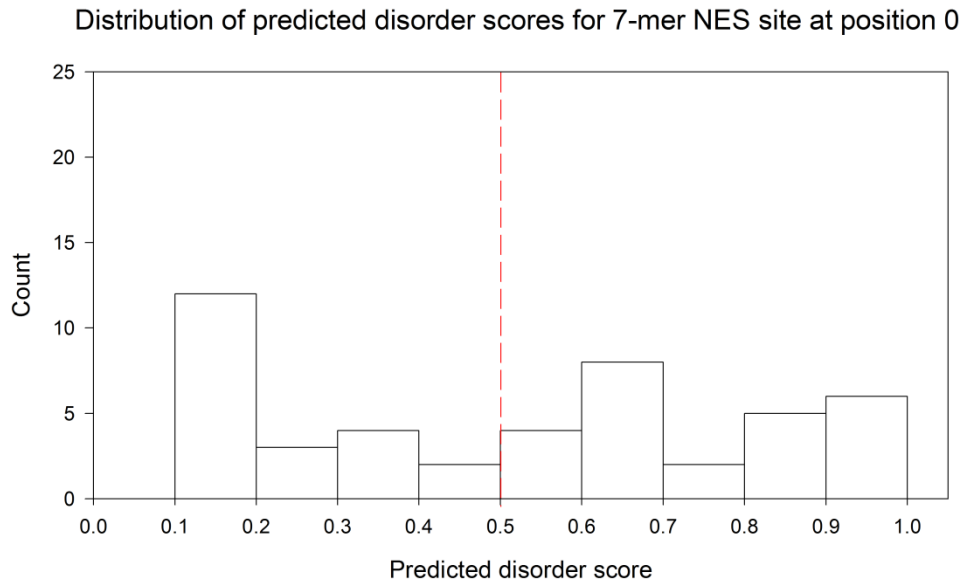




Figure 3.2.2: Histograms of predicted disorder scores at the first position of 6-mer and 7-mer NES sites. The red dash line indicates the cutoff value of POODLE-L (0.5).

## 3.3 *Classification of NES-containing proteins vs. non-NES-containing proteins*

### 3.3.1 Overview

To better characterize the NES, discrimination between NES-containing proteins and non-NES-containing proteins is an important issue which has not yet been addressed. In practice, a reliable classifier for this task will be useful in searching for potential NES-containing proteins. In this section, we report an extensive evaluation of the effectiveness of current methods for this task by different performance metrics. We also demonstrate and discuss the complete ineffectiveness of the consensus-based method for this task. Finally we discuss the high scoring background proteins.

### 3.3.2 The area under the receiver operating characteristic (ROC) curve

To evaluate current methods, we first applied each predictor to the mixed set of NES-containing test proteins and background proteins, and retrieved the highest predicted score for each protein to generate a ranked list (Figure 3.3.1). Based on this list, the performances were evaluated and compared by the receiver operating characteristic (ROC) curve and the area under the ROC curve (AUC). Meanwhile, the performances of current consensus sequences were also plotted in the ROC space by applying a simple rule that proteins which have a match to the given consensus are classified as "NES-containing". Surprisingly, Figure 3.3.2 shows that the corresponding points of the consensus sequences in ROC space are located below the diagonal, meaning that the performance is worse than random guessing. As for the computational methods, the AUC of flat NESsential (0.71) is higher than those of split NESsential (0.63) and NetNES (0.60), but none of them seem high enough for practical use. However, split NESsential provides some promising points, such as the point with 20% in true positive rate and 3% in false positive rate, which could be useful in searching for potential NES-containing proteins (Figure 3.3.2).

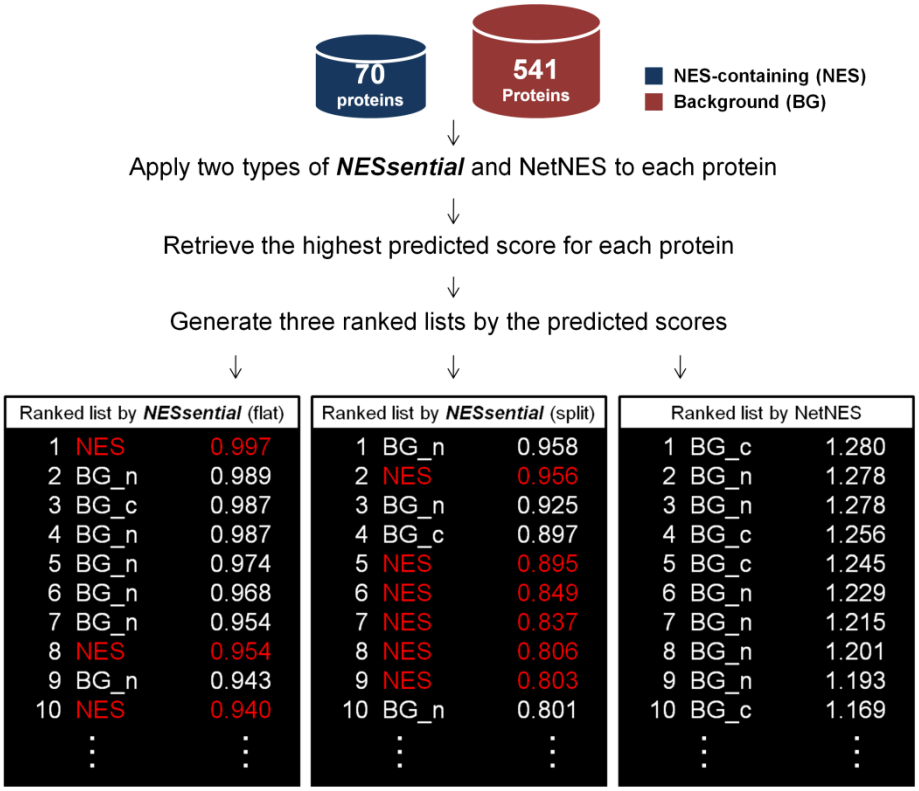



Figure 3.3.1: The pipeline of generating the ranked lists. NES, BG_n and BG_c in the ranked list denote NES-containing proteins, nuclear background proteins and cytosolic background proteins respectively.

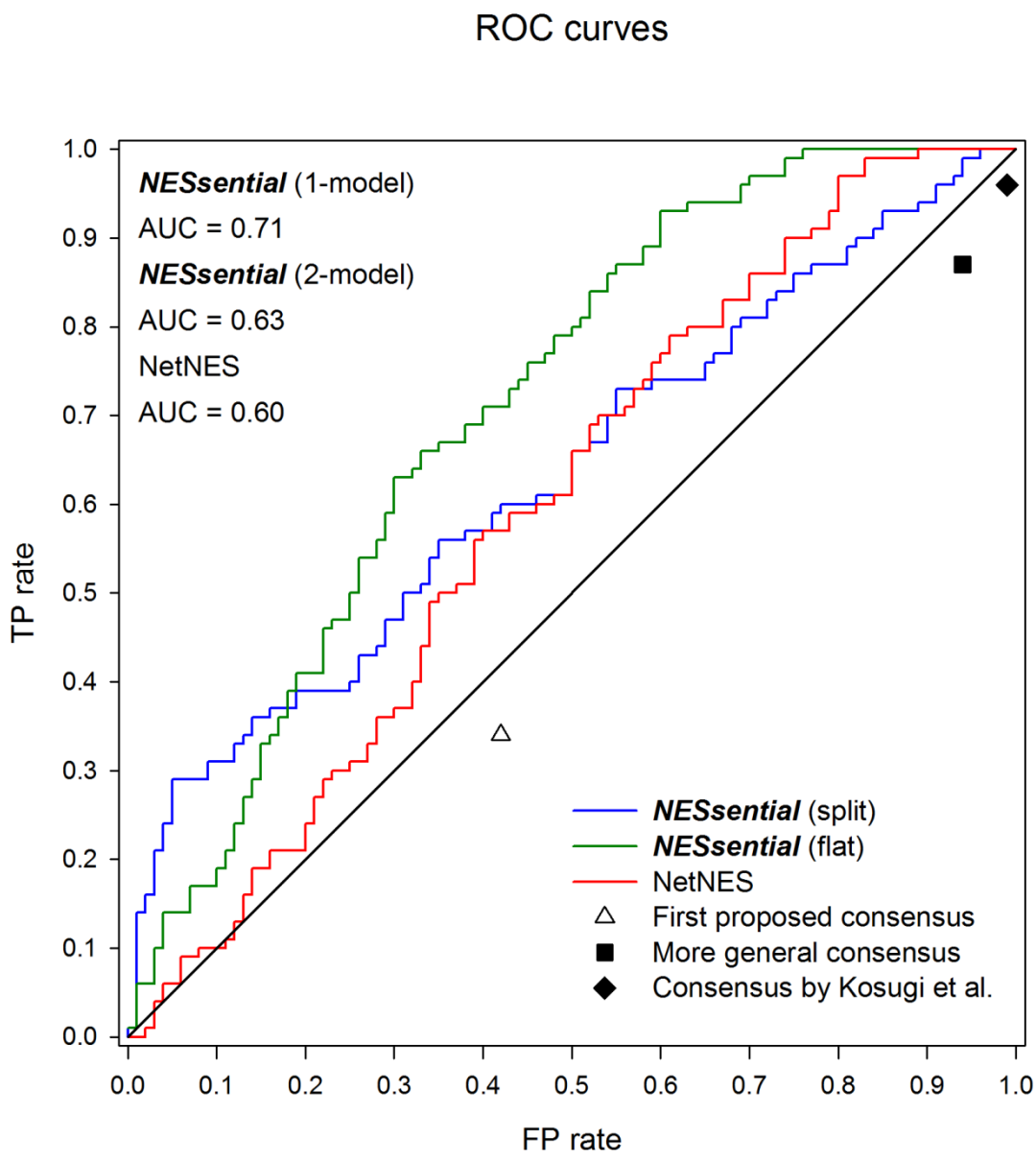



Figure 3.3.2: The ROC curves of two types of NESsential and NetNES. The dots denoting the performance of current consensus sequences are also plotted in the ROC space for comparison.

### 3.3.3 The precision-recall (PR) curve and the retrieval effectiveness

For further evaluation, we applied the Precision-Recall (PR) curve since it has been suggested that the PR curve can provide a more informative performance assessment than the ROC curve in the case of skewed datasets [19] as in our dataset. As measured by the 3- point and 11- point average precision, both flat and split NESsential demonstrate higher retrieval effectiveness than NetNES (Table 3.1). The PR-curve shown in Figure 3.3.3 further indicates that the difference between split NESsential and NetNES is mostly due to the much better precision in the low recall range, while flat NESsential attained higher precision than NetNES at all recall levels. Most significantly, split NESsential achieves a precision of over 0.5 at the recall points 0.1 and 0.2.

Table 3.1: The retrieval effectiveness among different methods

| | Flat NESsential | Split NESsential | NetNES |
|---|---|---|---|
| **3-point average precision** | 0.22 | 0.27 | 0.15 |
| **11-point average precision** | 0.28 | 0.31 | 0.23 |

*3-point average precision: the average precision at recall values of 20%, 50% and 80%.
*11-point average precision: the average precision at 11 standard recall points, from 0% through 100%.

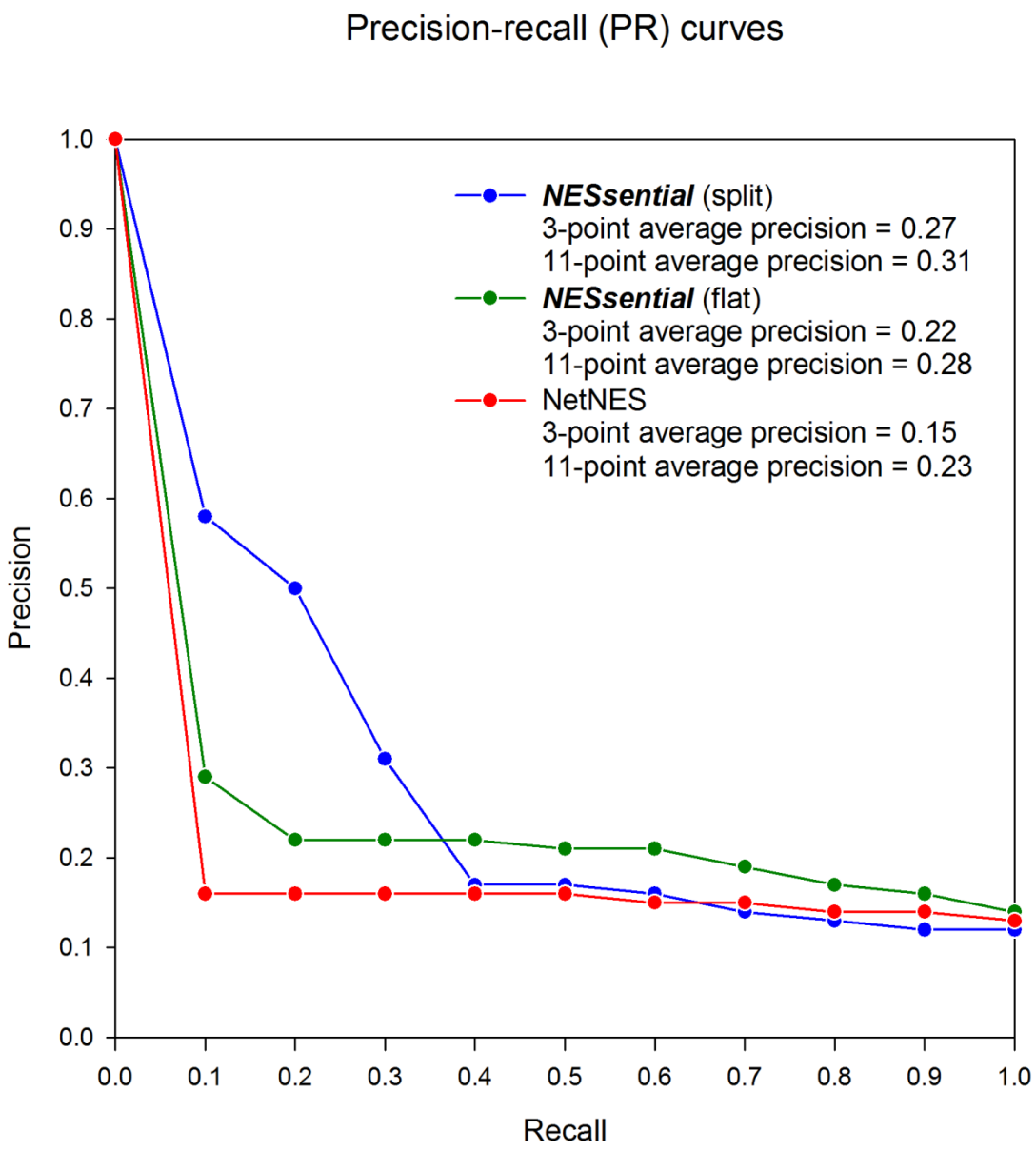



Figure 3.3.3: The precision-recall curves of two types of NESsential and NetNES.

### 3.3.4 Composition at the top positions of ranked lists

The stacked bar charts (Figure 3.3.4) provide another view showing the difference in composition at the higher positions in the ranked lists. Among the top ranked positions, flat and split NESsential list two to five times the number of NES-containing test proteins (dark grey stack in Figure 3.3.4) than NetNES. This result demonstrates that proteins with high scores predicted by NESsential, split NESsential especially, have a higher chance to contain NES's.

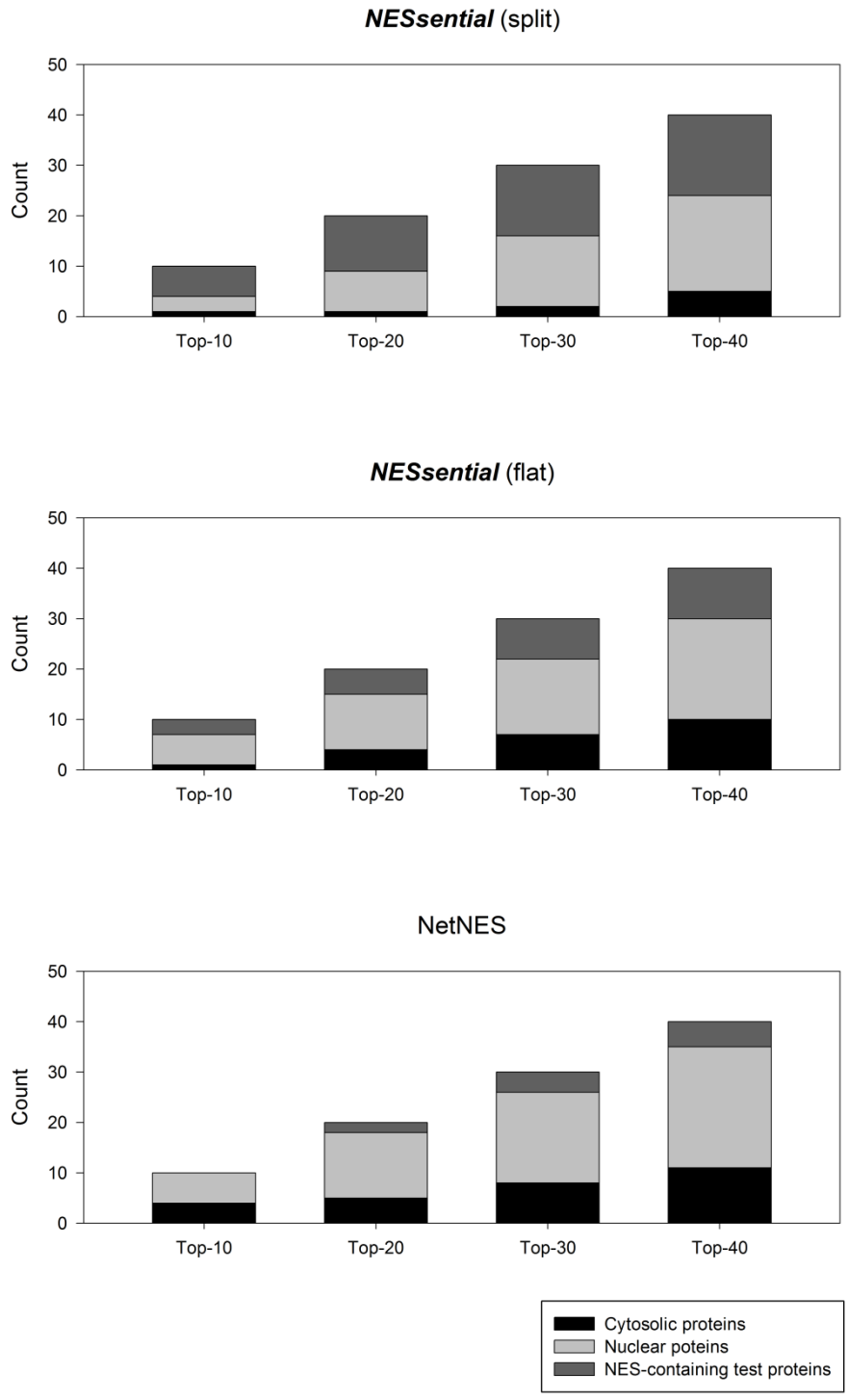



Figure 3.3.4: The stacked bar plots of composition at the top positions of the ranked lists made by two types of NESsential and NetNES.

### 3.3.5 Discussion

#### 3.3.5.1 The performance of consensus-based methods

The fact that both consensus sequences have negative predictive power indicates the "reversed" classification decision is more informative than the original one. In other words, it implies that background proteins are more likely to match the consensus sequences. To find a possible explanation, we calculated the expected number of occurrences for both consensus sequences in NES-containing, cytosolic and nuclear proteins respectively (we did not compute this for the Kosugi et al. "consensus", which is actually somewhat more complex than a simple regular expression). As the results show (Table 3.2), both consensus sequences are more likely to randomly occur in cytosolic and nuclear background proteins than in NES-containing proteins, due to differences in amino acid composition and average length. This can explain the negative correlation between matching the consensus sequences and whether a protein contains a NES.

Table 3.2: The expected numbers of two consensus sequences

| | Cytosolic | Nuclear | NES-containing |
|---|---|---|---|
| **The first proposed consensus matches** | 0.21 | 0.19 | 0.17 |
| **The more general consensus matches** | 3.9 | 2.8 | 2.5 |

*The first proposed consensus: L-x-(2,3)-[LIVFM]-x(2,3)-L-x-[LI].
*The more general consensus: [LIVFM]-x-(2,3)-[LIVFM]-x(2,3)-[LIVFM]-x-[LIVFM].

#### 3.3.5.2 Searching for potential novel NES-containing proteins

According to the PR curve, split NESsential is capable of retrieving 20% of NES-containing proteins with a precision of over 50%. Moreover, 6 out of 11 NES-containing proteins in the top-20 positions are correctly predicted not only at the protein level but also at the site level. These results demonstrate that proteins attaining a high split NESsential score have a high probability of containing NES's, and should be useful when searching for potential candidates. We, therefore, retrained NESsential on the all of the data (using training and test) and computed the scores for a set of nucleocytoplasmic dually localized yeast proteins (Table S2) downloaded from UniProt. Interestingly, one of the top-ranked proteins, the yeast nucleosome assembly protein (NAP1), was previously suggested to be exported by multiple proteins and CRM-1 might be one of its nuclear exporters [20] [21].

However, it should be mentioned that the current annotation of subcellular localization is not completely perfect, which means some of the cytosolic and nuclear proteins may contain undiscovered NES's, though the ratio is probably lower than for proteins annotated as dually localized. Therefore we also provide lists of nuclear (Table S3) and cytosolic proteins (Table S4) ranked by their scores given by NESsential (trained on all data). The top-ranked nuclear protein, the nitrogen regulatory protein GLN3 for instance, has been reported to contain a CRM1-mediated NES [22] (although not verified with leptomycin).

## 3.4 *Finding correct NES positions within NES-containing proteins*

### 3.4.1 Overview

In this section, we focus on the prediction task of finding the correct NES positions within NES-containing proteins as previously addressed by NetNES. Due to several complications, the different forms of prediction for instance, it's challenging to make a completely fair and objective comparison with NetNES. We first explicitly explain the complications and how they affected the evaluation and then report the performance of NESsential and NetNES.

### 3.4.2 Complications to making a fair comparison

To evaluate this prediction task previously addressed by NetNES, we plotted ROC curve as in the NetNES paper to estimate the residue-level prediction against 70 independent NES-containing proteins. However, unlike the protein-level classification task, there are some complications to making the comparison completely fair and objective. The first complication is the different forms of prediction between the two methods (Figure 3.4.1A). A conversion is required since NESsential makes predictions for each match of the pre-filter, while NetNES makes one for each residue. Figure 3.4.1A also illustrates the procedures of assigning the "site-level" predicted scores of NESsential to each residue of the sequence. Although the performances were comparable after the conversion, we note that the conversion yields a large amount of negative data -- the remaining residues without annotation of NES. This results in a highly skewed ratio of 1:40 between positive and negative data. There is another complication in ranking the predictions by their scores

to produce a PR-curve. This complication is caused by the different sources of gold standard data (Figure 3.4.1B). Some of the experimental data were verified by long deletions giving no exact boundaries of NES sites, yielding a mean length of 13.1 ± 9.8 residues which is much longer than the length of pre-filter. Due to this complication, 41% of the positive data, i.e. residues with annotation of NES, do not overlap with the pre-filter matches, and are therefore assigned a zero score by NESsential. On the other hand, though NetNES could make a positive prediction at any position theoretically, in fact it assigns a zero score to 61% of the positive data. The fact that such a large proportion of the positive data is assigned a score zero by both approaches makes it difficult to generate the rankings required for plotting complete PR-curves (Figure 3.4.2). For this reason, we used the ROC curve for assessment, despite the highly skewed ratio between positive and negative data.

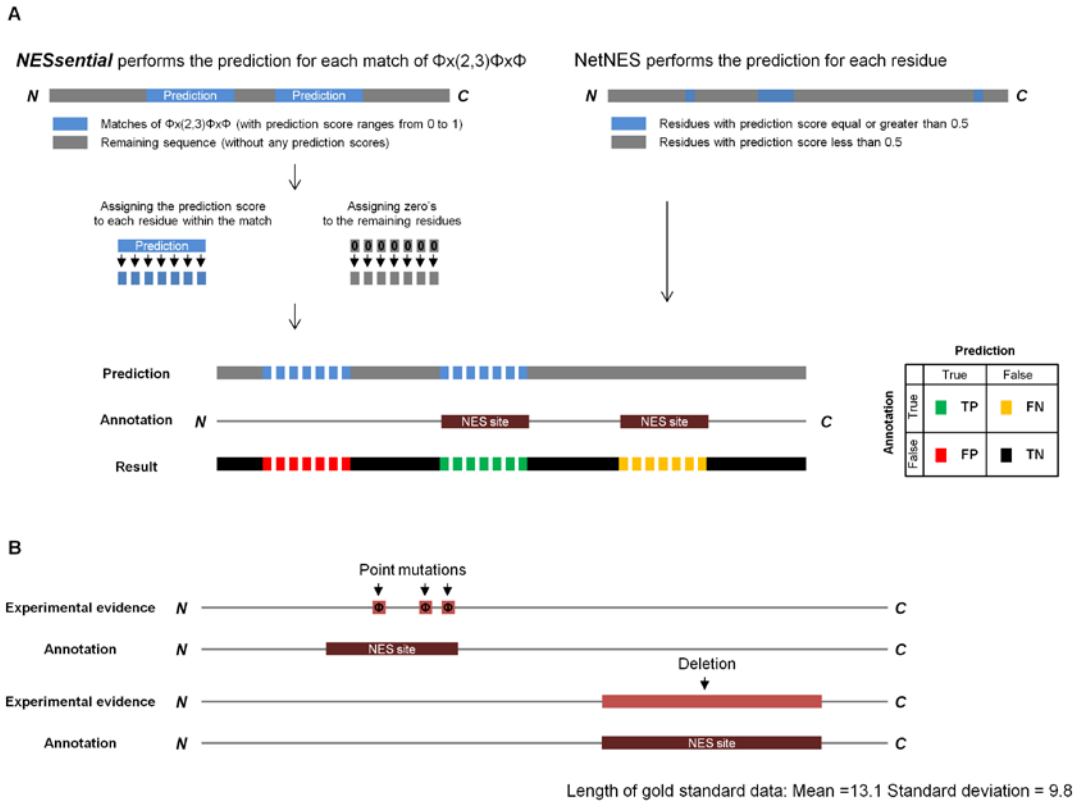



Figure 3.4.1: Two complications to making a fair comparison (A) The complication caused by different forms of prediction and the required conversion for comparison to NetNES. (B) The complication caused by different sources of gold standard data. Neither gives exact boundaries of NES sites.

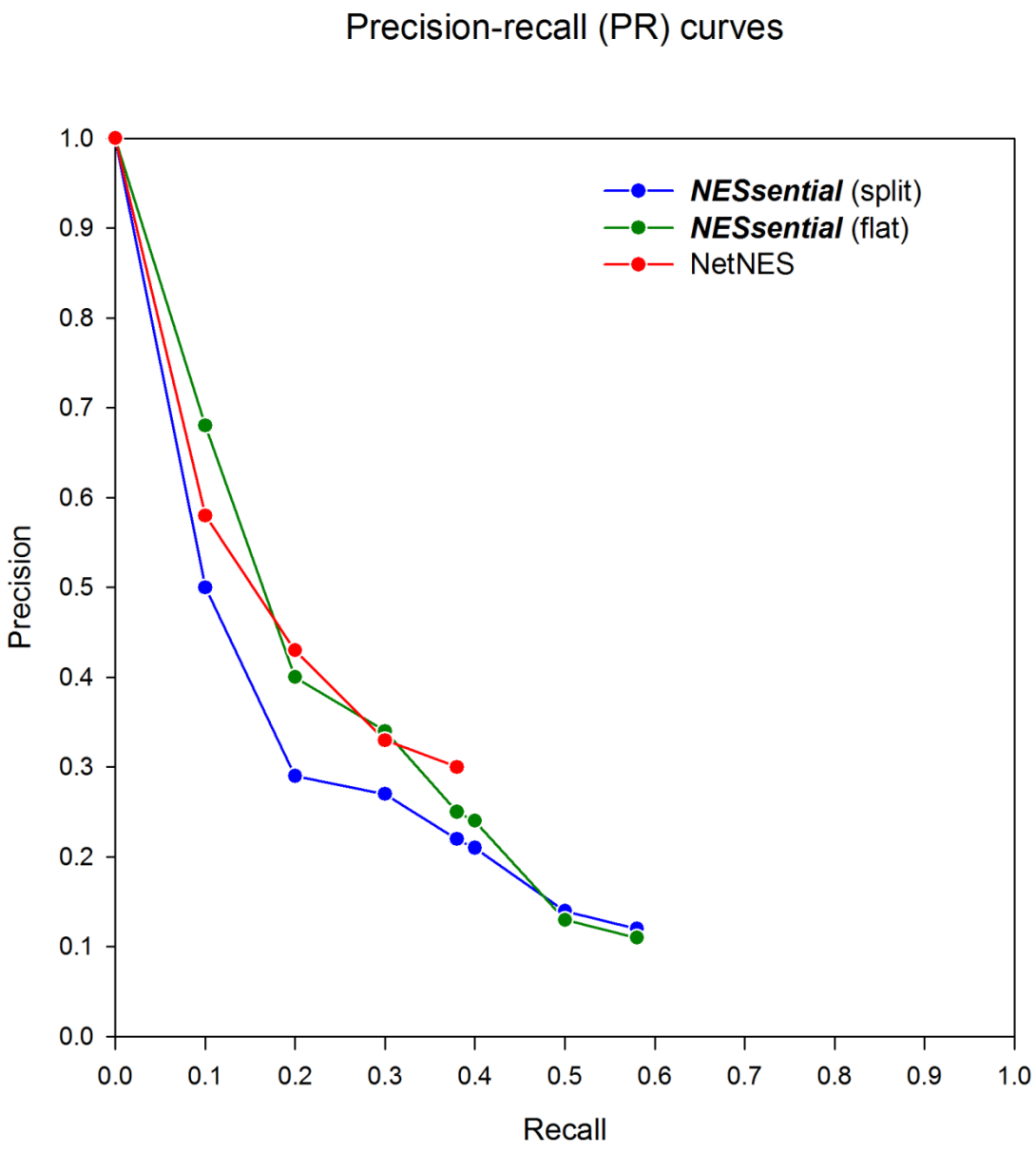



Figure 3.4.2: The incomplete precision-recall curves of two types of NESsential and NetNES.

### 3.4.3 The area under the receiver operating characteristic (ROC) curve

According to the incomplete PR-curves (Figure 3.4.2), the residue-level performance of flat NESsential and NetNES are better than split NESsential, especially at the low recall points. To make further comparison between flat NESsential and NetNES, we plotted ROC curves to measure the performance. Figure 3.4.3 shows the ROC curves of flat NESsential and NetNES. Before calculating the AUC, two dashed lines in ROC space should be first mentioned. As previously mentioned, a large proportion of the positive data was assigned "zero" by both predictors, causing a big jump in measurable performance. Consider the green dash line connecting the points (0.13, 0.58) and (1, 1) for instance, the point (0.13, 0.58) represents for the pair of false positive rate (FP rate) and true positive rate (TP rate) obtained by using the smallest non-zero score predicted by NESsential as a cutoff value, while the point (1, 1) represents for the unconditional assignment of all residues as NES positions. The AUC was calculated for each curve including the dashed line. As a result, flat NESsential achieved a higher AUC (0.75) than that of NetNES (0.68).

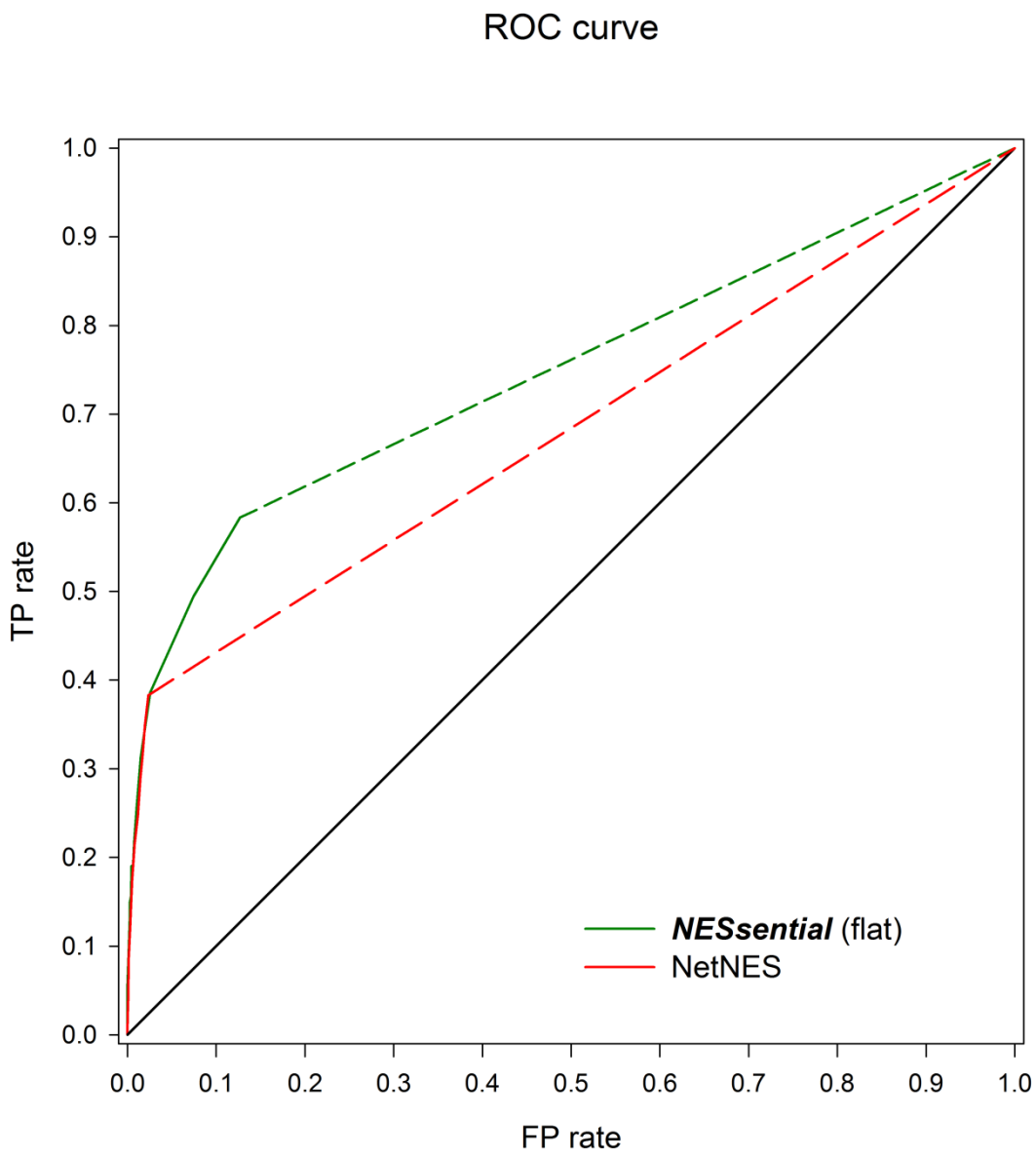



Figure 3.4.3: The ROC curves of flat NESsential and NetNES at the residue level.

### 3.4.4 Evaluating the site-level coverage

The classification accuracy at residue level is important, but for some tasks site level coverage is a more relevant measure. If one knows that a protein contains an NES (from leptomycin B inhibition, for example), but does not know the location of the NES site(s), one would like a tool which can predict those sites with an acceptable number of false predictions. As discussed in previous sections, a completely fair assessment of residue or site level prediction is not easy. Thus we developed another assessment involving coverage at the site level. We first applied flat NESsential, split NESsential and NetNES to generate three scores for each residue of the 70 test proteins (85 NES sites) as in the previous section. However, the longer length of prediction made by NESsential may give an unfair higher probability to overlap with NES sites. To exclude this effect, we only assigned the site score to the middle hydrophobic residue of the matching pre-filter and zero to all other residues.

By this conversion, each protein will have three lists of residues ranked by different predictors. The performance is then assessed by the site-level coverage, i.e. the proportion of NES sites overlapped by the top-N ranked predictions. Table 3.3 shows the site-level coverage by accepting up to the top-N ranked predictions. Flat and split NESsential perform better than NetNES after accepting the top-3 ranked predictions, and successfully locate approximately 73% of the test NES sites using four predictions per protein, while ten are needed by NetNES.

Table 3.3: The site-level coverage when accepting the top-N ranked predictions

| Top-N | Flat NESsential | Split NESsential | NetNES |
|---|---|---|---|
| **1** | 41% (35/85) | 38% (32/85) | 43% (36/85) |
| **2** | 56% (48/85) | 53% (45/85) | 55% (47/85) |
| **3** | 67% (57/85) | 65% (55/85) | 65% (55/85) |
| **4** | 73% (62/85) | 72% (61/85) | 67% (57/85) |
| **5** | 78% (66/85) | 79% (67/85) | 68% (58/85) |
| **6** | 79% (67/85) | 81% (69/85) | 71% (60/85) |
| **7** | 81% (69/85) | 82% (70/85) | 71% (60/85) |
| **8** | 86% (73/85) | 84% (71/85) | 72% (61/85) |
| **9** | 86% (73/85) | 85% (72/85) | 72% (61/85) |
| **10** | 88% (75/85) | 86% (73/85) | 73% (62/85) |

### 3.4.5 Discussion

Due to the various complications described above, it is challenging to make a completely fair comparison by either PR curve or ROC curve. Therefore in this section we provide two kinds of comparison. In residue level prediction, flat NESsential attained an AUC of 0.75, 0.07 higher than NetNES. We also developed an assessment of site-level coverage, a straightforward measurement indicating how well these predictors facilitate finding the correct position of NES's within NES-containing proteins. The results show that both types of NESsential can achieve 73% site-level coverage considering only the top 4 predictions for each protein, while the top 10 predictions are required for NetNES.

## 3.5 *Features and SVM models*

### 3.5.1 Cross-validation performance of SVM models

All SVM models used in NESsential were evaluated by the AUC under a 5-fold cross validation scheme. The SVM model used in flat NESsential achieved an AUC of 0.83, while the SVM models of the split NESsential for disordered and ordered pre-filter matches achieved an AUC of 0.92 and 0.75 respectively. We also applied another classification method, random forest and J48 decision tree to our training data. Table 3.4 shows that

models generated by SVM (implemented by libsvm-2.9) generally achieve higher AUC than by other classification method (implemented by Weka 3.6.2). Unfortunately, the cross validation scheme could not be directly applied to NetNES because we could not retrain it.

Table 3.4: The AUC values under 5-fold cross validation scheme

| | Flat NESsential | Split NESsential (disordered model) | Split NESsential (ordered model) |
|---|---|---|---|
| **SVM** | 0.83 | 0.92 | 0.75 |
| **Random forest** | 0.80 | 0.84 | 0.71 |
| **J48 decision tree** | 0.67 | 0.73 | 0.54 |

In this study, we reserved the test data for evaluation only when comparing to NetNES. However the combined dataset should give the best indication of the general performance of NESsential. Therefore we performed cross-validation on all of the data as well, resulting in AUCs of 0.80, 0.86 and 0.79, for the flat, split disordered and split ordered SVM models respectively.

### 3.5.2 Features used for prediction

Integrating new potentially relevant properties of NES function to those previously suggested, we extracted 22 biophysically inspired features from not only on the region matching the pre-filters, but also its upstream and downstream 10-mer flanks. These features mainly consist of (1) simple primary sequence attributes, such as the frequency of some specific amino acids: proline, negative-charged and polar residues, (2) predicted protein attributes (solvent accessibility and secondary structure by SABLE [23]; protein intrinsic disorder by Poodle-L), and (3) other properties such as the average hydrophobicity within the pre-filter matches (by the Kyte-Doolittle scale) and the distance in between the previous and next matches of pre-filters (or to the N- or C-terminal when no such match exists). We calculate flank disorder and solvent accessibility features based on a window of length 10, which requires special handling near the ends of sequences. For those matches close to the termini, we regard the "missing part" of such flanks as extremely disordered and accessible, assigning a disorder score of 1 and solvent accessibility of 100 to each missing "virtual residue". Table 3.5-3.7 provides a more detailed description of these 22 features.

To determine the relative importance of our features, we performed the F-score feature selection. The F-score is a commonly used statistical measure of the discriminative power for each feature by itself (see supplementary). The results indicated that the same features are used by the ordered and disordered SVM models of split NESsential, but with different relative importance (Table 3.5 and Table 3.6). The F-scores of the top features for disordered group are higher than those of the ordered group, which seems to be reflected in their respective AUC values. We also noticed that the top-ranked features for the disordered group are nearly all simple primary sequence attributes, while many of those for the ordered group are the predicted protein attributes, such as solvent accessibility and disorder. As for the model used in flat NESsential, the ranked list of features (Table 3.7) shows that the intrinsic disorder we proposed in this study is an important and relevant feature to NES-function. Since the F-score does not reveal mutual information among features, we also applied "leave-one-out" feature selection as a support. Although the differences in AUC after the removal of each feature are generally low, the combined analyses provide some further information about which top-ranked feature is more indispensable in the set of 22 features.

Table 3.5: The ranked list of features (disordered model of split NESsential)

| Rank | Feature description | F-score | ΔAUC |
|---|---|---|---|
| **1** | # of leucines among the 3 hydrophobic positions | 0.144 | 0.028 |
| **2** | # of negative-charged residues in the upstream flank | 0.102 | 0.053 |
| **3*** | # of polar residues in the downstream flank | 0.052 | 0.008 |
| **4** | Distance to previous match of ΦxΦxΦ divided by the protein length | 0.051 | 0.006 |
| **5** | Whether a hydrophobic residue exists in the upstream -4 position | 0.043 | -0.004 |
| **6*** | # of prolines within the pre-filter match ΦxΦxΦ | 0.043 | 0.005 |
| **7** | # of negative-charged residues in the downstream flank | 0.037 | 0.013 |
| **8** | # of negative-charged residues within the pre-filter match ΦxΦxΦ | 0.037 | 0.013 |
| **9*** | # of polar residues in the upstream flank | 0.031 | 0.003 |
| **10** | Avg. predicted solvent accessibility of downstream flank | 0.022 | -0.002 |
| **11*** | # of Methionines among the 3 hydrophobic positions | 0.013 | -0.008 |
| **12*** | Whether the first residue is involved in a beta-strand based on 2nd structure prediction | 0.008 | -0.003 |
| **13*** | Whether the first two residues are involved in a beta-strand based on 2nd structure prediction | 0.007 | -0.003 |
| **14*** | Expected number of pre-filter matches | 0.006 | 0.004 |
| **15** | Avg. predicted solvent accessibility of the upstream flank | 0.004 | 0.003 |
| **16** | Avg. predicted disorder score of the upstream flank | 0.003 | 0.007 |
| **17** | Difference of predicted solvent accessibilities (2nd and 3rd Φ position) | 0.001 | 0.002 |
| **18** | Distance to next match of ΦxΦxΦ divided by the protein length | 0.001 | -0.005 |
| **19** | Avg. predicted disorder score of the downstream flank | 0.001 | 0.009 |
| **20** | Avg. hydrophobicity of the pre-filter match ΦxΦxΦ | 0.000 | -0.006 |
| **21** | Avg. predicted disorder score of the pre-filter match ΦxΦxΦ | 0.000 | 0.004 |
| **22** | Avg. predicted solvent accessibility of the pre-filter match ΦxΦxΦ | 0.000 | 0.006 |

*indicates features that the mean value of spurious matches is greater than that of real NES site.

Table 3.6: The ranked list of features (ordered model of split NESsential)

| Rank | Feature description | F-score | ΔAUC |
|---|---|---|---|
| **1** | Whether a hydrophobic residue exists in the upstream -4 position | 0.038 | 0.015 |
| **2** | # of leucines among the 3 hydrophobic positions | 0.025 | -0.018 |
| **3** | # of negative-charged residues within the pre-filter match ΦxΦxΦ | 0.019 | 0.006 |
| **4** | Avg. predicted solvent accessibility of the pre-filter match ΦxΦxΦ | 0.014 | -0.014 |
| **5** | Avg. predicted disorder score of the pre-filter match ΦxΦxΦ | 0.012 | 0.021 |
| **6** | Avg. predicted disorder score of the downstream flank | 0.012 | 0.022 |
| **7** | # of polar residues in the downstream flank | 0.009 | -0.003 |
| **8*** | Whether the first two residues are involved in a beta-strand based on 2nd structure prediction | 0.006 | 0.004 |
| **9** | Avg. predicted disorder score of the upstream flank | 0.006 | -0.003 |
| **10** | Avg. predicted solvent accessibility of the downstream flank | 0.005 | -0.012 |
| **11*** | Avg. hydrophobicity of the pre-filter match ΦxΦxΦ | 0.005 | 0.003 |
| **12** | Expected number of pre-filter matches | 0.005 | -0.009 |
| **13** | Distance to previous match of ΦxΦxΦ divided by the protein length | 0.005 | -0.017 |
| **14** | # of prolines within the pre-filter match ΦxΦxΦ | 0.004 | 0.012 |
| **15*** | Whether the first residue is involved in a beta-strand based on 2nd structure prediction | 0.004 | 0.023 |
| **16** | Distance to next match of ΦxΦxΦ divided by the protein length | 0.003 | -0.002 |
| **17*** | # of negative-charged residues in the upstream flank | 0.002 | -0.019 |
| **18*** | Difference of predicted solvent accessibilities (2nd and 3rd Φ position) | 0.002 | -0.014 |
| **19*** | # of polar residues in the upstream flank | 0.001 | 0.008 |
| **20*** | # of negative-charged residues in the downstream flank | 0.000 | -0.007 |
| **21** | # of Methionines among the 3 hydrophobic positions | 0.000 | -0.019 |
| **22*** | Avg. predicted solvent accessibility of the upstream flank | 0.000 | -0.004 |

*indicates features that the mean value of spurious matches is greater than that of real NES site.

Table 3.7: The ranked list of features (model of flat NESsential)

| Rank | Feature description | F-score | ΔAUC |
|---|---|---|---|
| **1** | # of leucines among the 3 hydrophobic positions | 0.038 | 0.015 |
| **2** | Avg. predicted disorder score of the pre-filter match ΦxΦ | 0.025 | -0.018 |
| **3** | Avg. predicted disorder score of the downstream flank | 0.019 | 0.006 |
| **4** | Avg. predicted disorder score of the upstream flank | 0.014 | -0.014 |
| **5** | Whether a hydrophobic residue exists in the upstream -4 positions | 0.012 | 0.021 |
| **6** | # of negative-charged residues in the upstream flank | 0.012 | 0.022 |
| **7** | Distance to previous match of ΦxΦ divided by the protein length | 0.009 | -0.003 |
| **8** | # of negative-charged residues within the pre-filter match ΦxΦ | 0.006 | 0.004 |
| **9** | Avg. predicted solvent accessibility of the downstream flank | 0.006 | -0.003 |
| **10** | Avg. predicted solvent accessibility of the pre-filter match ΦxΦ | 0.005 | -0.012 |
| **11** | # of negative-charged residues in the downstream flank | 0.005 | 0.003 |
| **12*** | Whether the first two residues are involved in a beta-strand based on 2nd structure prediction | 0.005 | -0.009 |
| **13*** | Whether the first residue is involved in a beta-strand based on 2nd structure prediction | 0.005 | -0.017 |
| **14*** | # of prolines within the pre-filter match ΦxΦ | 0.004 | 0.012 |
| **15*** | Avg. hydrophobicity of the pre-filter match ΦxΦ | 0.004 | 0.023 |
| **16** | Avg. predicted solvent accessibility of the upstream flank | 0.003 | -0.002 |
| **17** | Distance to next match of ΦxΦ divided by the protein length | 0.002 | -0.019 |
| **18*** | # of polar residues in the upstream flank | 0.002 | -0.014 |
| **19** | Expected number of pre-filter matches | 0.001 | 0.008 |
| **20*** | Difference of predicted solvent accessibilities (2nd and 3rd Φ position) | 0.000 | -0.007 |
| **21*** | # of Methionines among the 3 hydrophobic positions | 0.000 | -0.019 |
| **22*** | # of polar residues in the downstream flank | 0.000 | -0.004 |

*indicates features that the mean value of spurious matches is greater than that of real NES site.

### 3.5.3 Discussion

#### 3.5.3.1 Sequence conservation as a relevant feature

Sequence conservation among orthologous proteins might be expected to provide useful information to improve NES recognition since the CRM1-mediated export pathway and the leucine-rich NES are found in all major branches of the eukaryotes. However, the spurious matches are often located in the hydrophobic core where the sequence is also conserved among orthologues. In fact, one should be careful when trying to apply sequence conservation to NES prediction, since NES's are not necessarily conserved among all orthologues. For example, the NES of the Snail transcription factor were found to be conserved only in mammalian orthologues while the NES is not present in other family members [24]. Another example indicates that the real NES of Human TPP1 is conserved among human, mice and frogs. However, the spurious matches of Human TPP1 also show high degrees of conservation between mammals [25]. Though these examples might be special cases, they show that NES's are not necessarily conserved among all orthologue families and a proper set of orthologous proteins is another issue requiring consideration.

#### 3.5.3.2 Directions for future improvement

Since the same feature set was used in training all SVM models in this study, it's interesting to discuss what caused the different performance in 5-fold cross-validation between the disordered and ordered models of split NESsential. One might speculate that the difference is a result of the different ratio between positive and negative data between the ordered and disordered pre-filter matches (see Figure 2.3). However, we tested this hypothesis by training models for the ordered group using randomly selected negative data to mimic the ratio found in the disordered group, but no significant improvement was observed. Thus it appears that the effect of unbalanced datasets cannot explain the difference in AUC, but rather the ordered NES's are less well described by our feature set. Our features mainly focus on the local information surrounding the NES site. However, the ordered NES's might be located in more buried regions and therefore require more complicated conformational changes to expose themselves to CRM1. Previous research has demonstrated some specific regulation, such as nearby phosphorylation sites [26] or the oligomeric state [27] of proteins with buried NES. Though these features seem to be required for specific proteins, we cannot exclude the possibility that these features will be

found in other NES-containing proteins and be helpful for future improvement, especially for the ordered group.

## 3.6 A case study of influenza viral proteins

In the previous section, we mentioned NES's may not always be conserved during evolution. In this regard, the non-structural (NS1) protein of influenza A viruses should be an excellent case to test how well NESsential and NetNES perform, as 100's of naturally occurring mutant variants are available. Recent research suggests that NES-mediated accumulation of NS1 in the cytoplasm increases the pathogenicity of the virus [28], which strengthens our expectation that the NES function of NS1 proteins should be conserved during evolution. We retrieved 327 full-length, non-identical NS1 protein sequences of human H1N1 influenza A viruses from the NCBI influenza virus resource [http://www.ncbi.nlm.nih.gov/genomes/FLU/FLU.html]. Within the dataset, 273 NS1 sequences are from the FLU project before 2009, while the remaining 54 sequences are from the 2009 pandemic. Figure 3.6.1 shows a phylogenetic tree of these sequences with the branch built by these pandemic 2009 and previous human H1N1 viral NS1 proteins.

We measured the prediction performance by site-level coverage (top-1 ranked) as described in section 3.4. As a result, split NESsential achieved a much higher site-level coverage (99%) than NetNES (61%). Interestingly, this difference was largely due to the site-level coverage of NS1 proteins from the 2009 pandemic (Table 3.7), which was reported to be more pathogenic than seasonal A (H1N1) virus [29][30]. Surprisingly, NetNES fails to detect many homologues even though they share high global sequence identity. The consistent protein-level coverage of NESsential suggests it may be more stable in predicting homologous NES's among NS1 proteins.

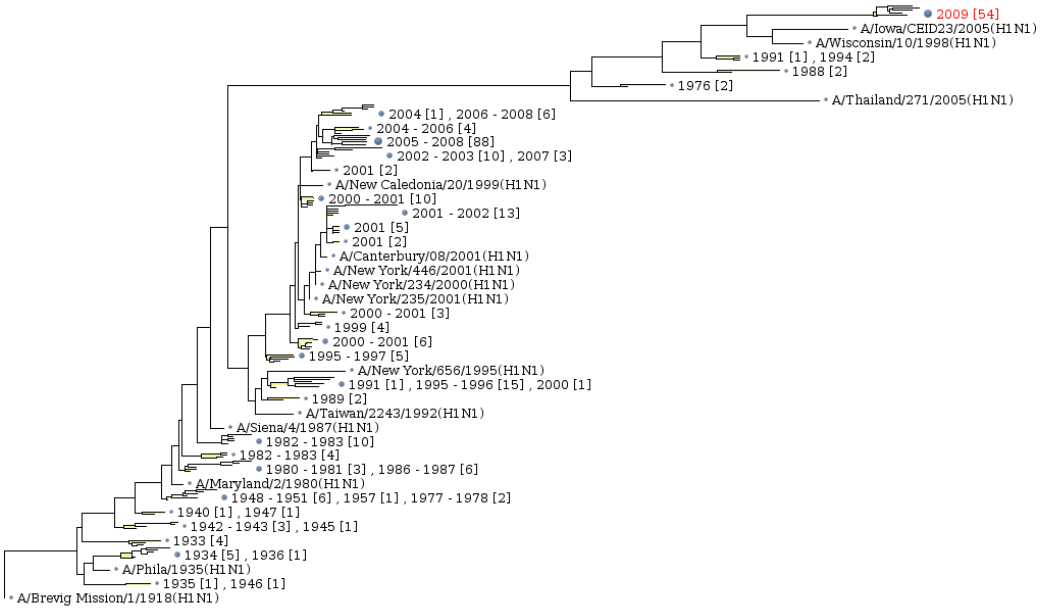



Figure 3.6.1: The phylogenetic tree provided by NCBI for the 2009 pandemic and previous human H1N1 viral NS1 proteins. The 2009 pandemic NS1 sequences are denoted in red with the number of NS1 proteins in parenthesis.

Table 3.7: The site-level coverage when accepting the top-1 ranked predictions

| | **Split NESsential** | **NetNES** |
|---|---|---|
| **Before 2009 (273)** | 99% | 73% |
| **2009 Pandemic (54)** | 100% | 2% |

# Chapter 4

# Conclusion

We have identified the different distribution of predicted protein disorder between real NES sites and spurious matches. Furthermore we have shown that the real NES sites can be naturally divided into two groups, ordered and disordered. Based on these analyses, we selected 22 biophysically inspired features and proposed NESsential, a SVM-based method implemented by LIBSVM. Meanwhile, we succeeded in enlarging the publicly available dataset of NES-containing proteins by about two-fold and used these independent and newly discovered proteins for evaluation. This up-to-date resource is a valuable resource for future analysis work.

For the task of classifying NES-containing proteins from non-NES-containing proteins, both flat and split NESsential achieved higher retrieval effectiveness than the state-of-the-art predictor, NetNES. Indeed the consensus sequence methods are completely ineffective for this task and NetNES also performs quite poorly. In contrast, split NESsential provides a practical precision of over 50% at low recall level, which should be useful in searching for potential candidates containing leucine-rich NES's. Besides, NESsential also achieved a higher AUC (by 0.07) and a higher site-level coverage than NetNES in the task of finding correct NES positions within NES-containing proteins.

# Reference

1. Diella F, Haslam N, Chica C, Budd A, Michael S, Brown NP, Trave G, Gibson TJ: **Understanding eukaryotic linear motifs and their role in cell signaling and regulation**. *Front. Biosci* 2008, **13**:6580-6603.

2. Turner JG, Sullivan DM: **CRM1-mediated nuclear export of proteins and drug resistance in cancer**. *Current Medicinal Chemistry* 2008, **15**:2648–2655.

1. Nilsen,T., Rosendal,K.R., Sorensen,V., Wesche,J., Olsnes,S. and Wiedlocha,A. (2007) A Nuclear Export Sequence Located on a beta-Strand in Fibroblast Growth Factor-1. *Journal of Biological Chemistry*, **282**, 26245-26256, 10.1074/jbc.M611234200.

4. Monecke T, Guttler T, Neumann P, Dickmanns A, Gorlich D, Ficner R: **Crystal Structure of the Nuclear Export Receptor CRM1 in Complex with Snurportin1 and RanGTP**. *Science* 2009, **324**:1087-1091.

5. Dong X, Biswas A, Süel KE, Jackson LK, Martinez R, Gu H, Chook YM: **Structural basis for leucine-rich nuclear export signal recognition by CRM1**. *Nature* 2009, **458**:1136-1141.

6. Bogerd HP, Fridell RA, Benson RE, Hua J, Cullen BR: **Protein sequence requirements for function of the human T-cell leukemia virus type 1 Rex nuclear export signal delineated by a novel in vivo randomization-selection assay.** *Mol Cell Biol* 1996, **16**:4207-4214.

7. Fischer U, Huber J, Boelens WC, Mattaj IW, Lührmann R: **The HIV-1 Rev activation domain is a nuclear export signal that accesses an export pathway used by specific cellular RNAs**. *Cell* 1995, **82**:475-483.

8. Wen W, Meinkoth JL, Tsien RY, Taylor SS: **Identification of a signal for rapid export of proteins from the nucleus**. *Cell* 1995, **82**:463-473.

9. la Cour T, Gupta R, Rapacki K, Skriver K, Poulsen FM, Brunak S: **NESbase version 1.0: a database of nuclear export signals**. *Nucleic Acids Res* 2003, **31**:393-396.

10. la Cour T, Kiemer L, Mølgaard A, Gupta R, Skriver K, Brunak S: **Analysis and prediction of leucine-rich nuclear export signals**. *Protein Engineering Design and Selection* 2004, **17**:527-536.

11. Kosugi S, Hasebe M, Tomita M, Yanagawa H: **Nuclear export signal consensus sequences defined using a localization-based yeast selection system**. *Traffic* 2008, **9**:2053-2062.

12. Dunker AK, Cortese MS, Romero P, Iakoucheva LM, Uversky VN: **Flexible nets. The roles of intrinsic disorder in protein interaction networks**. *FEBS J* 2005, **272**:5129-5148.

13. Uversky VN, Oldfield CJ, Dunker AK: **Showing your ID: intrinsic disorder as an ID for recognition, regulation and cell signaling**. *J. Mol. Recognit* 2005, **18**:343-384.

14. Kutay U, Güttinger S: **Leucine-rich nuclear-export signals: born to be weak**. *Trends in Cell Biology* 2005, **15**:121–124.

15. Hirose S, Shimizu K, Kanai S, Kuroda Y, Noguchi T: **POODLE-L: a two-level SVM prediction system for reliably predicting long disordered regions**. *Bioinformatics* 2007, **23**:2046-2053.

16. Ward JJ, McGuffin LJ, Bryson K, Buxton BF, Jones DT: **The DISOPRED server for the prediction of protein disorder**. *Bioinformatics* 2004, **20**:2138 -2139.

17. Chang C, Lin C: **LIBSVM: a Library for Support Vector Machines**. 2001.

18. Kudo N, Taoka H, Toda T, Yoshida M, Horinouchi S: **A novel nuclear export signal sensitive to oxidative stress in the fission yeast transcription factor Pap1**. *J. Biol. Chem* 1999, **274**:15151-15158.

19. He H, Garcia EA: **Learning from Imbalanced Data**. *IEEE Trans. Knowl. Data Eng.* 2009, **21**:1263-1284.

20. Mosammaparast N, Ewart CS, Pemberton LF: **A role for nucleosome assembly protein 1 in the nuclear transport of histones H2A and H2B**. *EMBO J* 2002, **21**:6527-6538.

21. Park Y, Luger K: **The structure of nucleosome assembly protein 1**. *Proc Natl Acad Sci U S A* 2006, **103**:1248-1253.

22. Carvalho J, Zheng XFS: **Domains of Gln3p interacting with karyopherins, Ure2p, and the target of rapamycin protein**. *J. Biol. Chem* 2003, **278**:16878-16886.

23. Adamczak R, Porollo A, Meller J: **Combining prediction of secondary structure and solvent accessibility in proteins**. *Proteins* 2005, **59**:467-475.

24. Garcia de Herreros A, Dominguez D, Montserrat-Sentis B, Virgos-Soler A, Guaita S, Grueso J, Porta M, Puig I, Baulida J, Franci C: **Phosphorylation regulates the subcellular location and activity of the snail transcriptional repressor**. *Molecular and Cellular Biology* 2003, **23**:5078.

25. Chen L, Liu D, Songyang Z: **Telomere Maintenance through Spatial Control of Telomeric Proteins**. *Molecular and Cellular Biology* 2007, **27**:5898-5909.

26. Meng W, Swenson LL, Fitzgibbon MJ, Hayakawa K, Ter Haar E, Behrens AE, Fulghum JR, Lippke JA: **Structure of mitogen-activated protein kinase-activated protein (MAPKAP) kinase 2 suggests a bifunctional switch that couples kinase activation with nuclear export**. *J. Biol. Chem* 2002, **277**:37401-37405.

27. Stommel JM, Marchenko ND, Jimenez GS, Moll UM, Hope TJ, Wahl GM: **A leucine-rich nuclear export signal in the p53 tetramerization domain: regulation of subcellular localization and p53 activity by NES masking**. *EMBO J* 1999, **18**:1660-1672.

28. Keiner B, Maenz B, Wagner R, Cattoli G, Capua I, Klenk H: **Intracellular distribution of NS1 correlates with the infectivity and interferon antagonism of an avian influenza virus (H7N1)**. *J. Virol* 2010, **84**:11858-11865.

29. Maines TR, Jayaraman A, Belser JA, Wadford DA, Pappas C, Zeng H, Gustin KM, Pearce MB, Viswanathan K, Shriver ZH, Raman R, Cox NJ, Sasisekharan R, Katz JM, Tumpey TM: **Transmission and pathogenesis of swine-origin 2009 A(H1N1) influenza viruses in ferrets and mice**. *Science* 2009, **325**:484-487.

30. Munster VJ, de Wit E, van den Brand JMA, Herfst S, Schrauwen EJA, Bestebroer TM, van de Vijver D, Boucher CA, Koopmans M, Rimmelzwaan GF, Kuiken T, Osterhaus ADME, Fouchier RAM: **Pathogenesis and transmission of swine-origin 2009 A(H1N1) influenza virus in ferrets**. *Science* 2009, **325**:481-483.

31. Hood J, Hwang W, Silver P: **The Saccharomyces cerevisiae cyclin Clb2p is targeted to multiple subcellular locations by cis- and trans-acting determinants**. *J Cell Sci* 2001, **114**:589-597.

32. Hood-DeGrenier JK, Boulton CN, Lyo V: **Cytoplasmic Clb2 is required for timely inactivation of the mitotic inhibitor Swe1 and normal bud morphogenesis in Saccharomyces cerevisiae**. *Curr Genet* 2006, **51**:1-18.

33. Suetsugu S: **Translocation of N-WASP by Nuclear Localization and Export Signals into the Nucleus Modulates Expression of HSP90**. *Journal of Biological Chemistry* 2003, **278**:42515-42523.

34. Bembenek J, Kang J, Kurischko C, Li B, Raab JR, Belanger KD, Luca FC, Yu H: **Crm1-mediated nuclear export of Cdc14 is required for the completion of cytokinesis in budding yeast**. *Cell Cycle* 2005, **4**:961-971.

35. Colnaghi R, Connell CM, Barrett RMA, Wheatley SP: **Separating the Anti-apoptotic and Mitotic Roles of Survivin**. *Journal of Biological Chemistry* 2006, **281**:33450-33456.

36. Knauer SK, Bier C, Habtemichael N, Stauber RH: **The Survivin–Crm1 interaction is essential for chromosomal passenger complex localization and function**. *EMBO Rep* 2006, **7**:1259-1265.

37. Stauber RH, Rabenhorst U, Rekik A, Engels K, Bier C, Knauer SK:

**Nucleocytoplasmic Shuttling and the Biological Activity of Mouse Survivin are Regulated by an Active Nuclear Export Signal**. *Traffic* 2006, **7**:1461-1472.

38. Kang WK, Kurihara M, Matsumoto S: **The BRO proteins of Bombyx mori nucleopolyhedrovirus are nucleocytoplasmic shuttling proteins that utilize the CRM1-mediated nuclear export pathway**. *Virology* 2006, **350**:184–191.

39. Tong EHY: **Regulation of Nucleocytoplasmic Trafficking of Transcription Factor OREBP/TonEBP/NFAT5**. *Journal of Biological Chemistry* 2006, **281**:23870-23879.

40. Akoumianaki T, Kardassis D, Polioudaki H, Georgatos SD, Theodoropoulos PA: **Nucleocytoplasmic shuttling of soluble tubulin in mammalian cells**. *J. Cell. Sci* 2009, **122**:1111-1118.

41. Chevalier SA, Meertens L, Calattini S, Gessain A, Kiemer L, Mahieux R: **Presence of a functional but dispensable nuclear export signal in the HTLV-2 Tax protein**. *Retrovirology* 2005, **2**:70.

42. Falini B, Bolli N, Shan J, Martelli MP, Liso A, Pucciarini A, Bigerna B, Pasqualucci L, Mannucci R, Rosati R, others: **Both carboxy-terminus NES motif and mutated tryptophan (s) are crucial for aberrant nuclear export of nucleophosmin leukemic mutants in NPMc+ AML**. *Blood* 2006, **107**:4514.

43. Mariano AR, Colombo E, Luzi L, Martinelli P, Volorio S, Bernard L, Meani N, Bergomas R, Alcalay M, Pelicci PG: **Cytoplasmic localization of NPM in myeloid leukemias is dictated by gain-of-function mutations that create a functional nuclear export signal**. *Oncogene* 2006, **25**:4376–4380.

44. Wang W, Budhu A, Forgues M, Wang XW: **Temporal and spatial control of nucleophosmin by the Ran–Crm1 complex in centrosome duplication**. *Nat Cell Biol* 2005, **7**:823-830.

45. Cheng G, Brett ME, He B: **Signals That Dictate Nuclear, Nucleolar, and Cytoplasmic Shuttling of the γ134.5 Protein of Herpes Simplex Virus Type 1**.

*Journal of virology* 2002, **76**:9434.

46. Eulálio A, Nunes-Correia I, Carvalho AL, Faro C, Citovsky V, Salas J, Salas ML, Simões S, de Lima MC: **Nuclear export of African swine fever virus p37 protein occurs through two distinct pathways and is mediated by three independent signals**. *Journal of virology* 2006, **80**:1393.

47. Tsukahara F: **Identification of Novel Nuclear Export and Nuclear Localization-related Signals in Human Heat Shock Cognate Protein 70**. *Journal of Biological Chemistry* 2004, **279**:8867-8872.

48. Turner JG, Engel R, Derderian JA, Jove R, Sullivan DM: **Human topoisomerase IIalpha nuclear export is mediated by two CRM-1-dependent nuclear export signals**. *J. Cell. Sci* 2004, **117**:3061-3071.

49. Mirski SE, Bielawski JC, Cole SP: **Identification of functional nuclear export sequences in human topoisomerase II alpha and beta**. *Biochemical and Biophysical Research Communications* 2003, **306**:905–911.

50. Liu H, Deng X, Shyu YJ, Li JJ, Taparowsky EJ, Hu CD: **Mutual regulation of c-Jun and ATF2 by transcriptional activation and subcellular localization**. *The EMBO journal* 2006, **25**:1058–1069.

51. Davidson PJ, Li SY, Lohse AG, Vandergaast R, Verde E, Pearson A, Patterson RJ, Wang JL, Arnoys EJ: **Transport of galectin-3 between the nucleus and cytoplasm. I. Conditions and signals for nuclear import**. *Glycobiology* 2006, **16**:602.

52. Lischka P, Rauh C, Mueller R, Stamminger T: **Human Cytomegalovirus UL84 Protein Contains Two Nuclear Export Signals and Shuttles between the Nucleus and the Cytoplasm**. *Journal of Virology* 2006, **80**:10274-10280.

53. Yoshida H, Oku M, Suzuki M, Mori K: **pXBP1 (U) encoded in XBP1 pre-mRNA negatively regulates unfolded protein response activator pXBP1 (S) in mammalian ER stress response**. *The Journal of cell biology* 2006, **172**:565.

54. Singhal PK, Rajendra Kumar P, Subba Rao MRK, Mahalingam S: **Nuclear Export of Simian Immunodeficiency Virus Vpx Protein**. *Journal of Virology* 2006, **80**:12271-12282.

55. Makita J, Kurooka H, Mori K, Akagi Y, Yokota Y: **Identification of the nuclear export signal in the helix-loop-helix inhibitor Id1**. *FEBS letters* 2006, **580**:1812–1816.

56. Nishiyama K, Takaji K, Uchijima Y, Kurihara Y, Asano T, Yoshimura M, Ogawa H, Kurihara H: **Protein Kinase A-regulated Nucleocytoplasmic Shuttling of Id1 during Angiogenesis**. *Journal of Biological Chemistry* 2007, **282**:17200-17209.

57. Baranek C, Sock E, Wegner M: **The POU protein Oct-6 is a nucleocytoplasmic shuttling protein**. *Nucleic acids research* 2005, **33**:6277.

58. Pasdeloup D, Poisson N, Raux H, Gaudin Y, Ruigrok RW, Blondel D: **Nucleocytoplasmic shuttling of the rabies virus P protein requires a nuclear localization signal and a CRM1-dependent nuclear export signal**. *Virology* 2005, **334**:284–293.

59. Kong KY: **Cytoplasmic Nuclear Transfer of the Actin-capping Protein Tropomodulin**. *Journal of Biological Chemistry* 2004, **279**:30856-30864.

60. Verhagen J, Donnelly M, Elliott G: **Characterization of a Novel Transferable CRM-1-Independent Nuclear Export Signal in a Herpesvirus Tegument Protein That Shuttles between the Nucleus and Cytoplasm**. *Journal of Virology* 2006, **80**:10021-10035.

61. Chester A, Somasekaram A, Tzimina M, Jarmuz A, Gisbourne J, O'Keefe R, Scott J, Navaratnam N: **The apolipoprotein B mRNA editing complex performs a multifunctional cycle and suppresses nonsense-mediated decay**. *The EMBO Journal* 2003, **22**:3971–3982.

62. Hwang CY, Kim IY, Kwon K: **Cytoplasmic localization and ubiquitination of p21(Cip1) by reactive oxygen species**. *Biochem. Biophys. Res. Commun* 2007, **358**:219-225.

63. Bachmann RA: **A Nuclear Transport Signal in Mammalian Target of Rapamycin Is Critical for Its Cytoplasmic Signaling to S6 Kinase 1**. *Journal of Biological Chemistry* 2006, **281**:7357-7363.

64. Xia J: **Huntingtin contains a highly conserved nuclear export signal**. *Human Molecular Genetics* 2003, **12**:1393-1403.

65. Maekawa M, Yamamoto T, Nishida E: **Regulation of subcellular localization of the antiproliferative protein Tob by its nuclear export signal and bipartite nuclear localization signal sequences**. *Experimental cell research* 2004, **295**:59–65.

66. Han X, Saito H, Miki Y, Nakanishi A: **A CRM1-mediated nuclear export signal governs cytoplasmic localization of BRCA2 and is essential for centrosomal localization of BRCA2**. *Oncogene* 2007, **27**:2969–2977.

67. Wilson JM, Le VQ, Zimmerman C, Marmorstein R, Pillus L: **Nuclear export modulates the cytoplasmic Sir2 homologue Hst2**. *EMBO Rep* 2006, **7**:1247-1251.

68. Moshynskyy I, Viswanathan S, Vasilenko N, Lobanov V, Petric M, Babiuk LA, Zakhartchouk AN: **Intracellular localization of the SARS coronavirus protein 9b: evidence of active export from the nucleus**. *Virus research* 2007, **127**:116–121.

69. Tsuchiya A, Tashiro E, Yoshida M, Imoto M: **Involvement of protein phosphatase 2A nuclear accumulation and subsequent inactivation of activator protein-1 in leptomycin B-inhibited cyclin D1 expression**. *Oncogene* 2006, **26**:1522–1532.

70. Honda T, Nakajima K: **Mouse Disabled1 (DAB1) Is a Nucleocytoplasmic Shuttling Protein**. *Journal of Biological Chemistry* 2006, **281**:38951-38965.

71. Dominguez D, Montserrat-Sentis B, Virgos-Soler A, Guaita S, Grueso J, Porta M, Puig I, Baulida J, Franci C, Garcia de Herreros A: **Phosphorylation regulates the subcellular location and activity of the snail transcriptional repressor**. *Molecular and cellular biology* 2003, **23**:5078.

72. Velichkova M, Hasson T: **Keap1 regulates the oxidation-sensitive shuttling of Nrf2 into and out of the nucleus via a Crm1-dependent nuclear export mechanism**. *Molecular and cellular biology* 2005, **25**:4501.

73. Husberg C, Murphy P, Bjørgo E, Kalland KH, Kolstø AB: **Cellular localisation and nuclear export of the human bZIP transcription factor TCF11**. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research* 2003, **1640**:143–151.

74. Wang Y, Kakinuma N, Zhu Y, Kiyama R: **Nucleo-cytoplasmic shuttling of human Kank protein accompanies intracellular translocation of beta-catenin**. .

75. Li W: **Nrf2 Possesses a Redox-insensitive Nuclear Export Signal Overlapping with the Leucine Zipper Motif**. *Journal of Biological Chemistry* 2005, **280**:28430-28438.

76. Li W: **Nrf2 Possesses a Redox-sensitive Nuclear Exporting Signal in the Neh5 Transactivation Domain**. *Journal of Biological Chemistry* 2006, **281**:27251-27263.

77. Hamuro J, Higuchi O, Okada K, Ueno M, Iemura S, Natsume T, Spearman H, Beeson D, Yamanashi Y: **Mutations Causing DOK7 Congenital Myasthenia Ablate Functional Motifs in Dok-7**. *Journal of Biological Chemistry* 2007, **283**:5518-5524.

78. Park SJ, Lee D, Choi CY, Ryu SY: **Induction of apoptosis by NORE1A in a manner dependent on its nuclear export**. *Biochemical and biophysical research communications* 2008, **368**:56–61.

79. Yanai H, Kobayashi T, Hayashi Y, Watanabe Y, Ohtaki N, Zhang G, De La Torre JC, Ikuta K, Tomonaga K: **A methionine-rich domain mediates CRM1-dependent nuclear export activity of Borna disease virus phosphoprotein**. *Journal of virology* 2006, **80**:1121.

80. Jain AK: **Nuclear Import and Export Signals in Control of Nrf2**. *Journal of Biological Chemistry* 2005, **280**:29158-29168.

81. Saito Y, Yamagishi N, Hatayama T: **Different localization of Hsp105 family**

**proteins in mammalian cells**. *Experimental cell research* 2007, **313**:3707–3717.

82. Miki T: **Alternative Splicing of Staufen2 Creates the Nuclear Export Signal for CRM1 (Exportin 1)**. *Journal of Biological Chemistry* 2004, **279**:47473-47479.

83. Zhang A, Li C, Tsai S, Chen JD: **Subcellular localization of ankyrin repeats cofactor-1 regulates its corepressor activity**. *J. Cell. Biochem.* 2007, **101**:1301-1315.

84. Ichikawa HT, Sowden MP, Torelli AT, Bachl J, Huang P, Dance GS, Marr SH, Robert J, Wedekind JE, Smith HC, others: **Structural phylogenetic analysis of activation-induced deaminase function**. *The Journal of Immunology* 2006, **177**:355.

85. Engelsma D, Valle N, Fish A, Salome N, Almendral JM, Fornerod M: **A supraphysiological nuclear export signal is required for parvovirus nuclear export**. *Molecular biology of the cell* 2008, **19**:2544.

86. North BJ, Verdin E: **Interphase nucleo-cytoplasmic shuttling and localization of SIRT2 during mitosis**. *PLoS One* 2007, **2**:784.

87. Liu L, Chen G, Ji X, Gao G: **ZAP is a CRM1-dependent nucleocytoplasmic shuttling protein**. *Biochemical and biophysical research communications* 2004, **321**:517–523.

88. Houlard M, Romero-Portillo F, Germani A, Depaux A, Regnier-Ricard F, Gisselbrecht S, Varin-Blank N: **Characterization of VIK-1: a new Vav-interacting Kruppel-like protein**. *Oncogene* 2004, **24**:28-38.

89. González-Mariscal L, Ponce A, Alarcón L, Jaramillo BE: **The tight junction protein ZO-2 has several functional nuclear export signals**. *Experimental cell research* 2006, **312**:3323–3335.

90. Shapiro MJ: **The Carboxyl-terminal Segment of the Adaptor Protein ALX Directs Its Nuclear Export during T Cell Activation**. *Journal of Biological Chemistry* 2005, **280**:38242-38246.

91. Papp LV, Lu J, Striebel F, Kennedy D, Holmgren A, Khanna KK: **The Redox State of SECIS Binding Protein 2 Controls Its Localization and Selenocysteine Incorporation Function**. *Molecular and Cellular Biology* 2006, **26**:4895-4910.

92. Heilman DW, Teodoro JG, Green MR: **Apoptin Nucleocytoplasmic Shuttling Is Required for Cell Type-Specific Localization, Apoptosis, and Recruitment of the Anaphase-Promoting Complex/Cyclosome to PML Bodies**. *Journal of Virology* 2006, **80**:7535-7545.

93. Munoz-Fontela C, Collado M, Rodriguez E, Garcia MA, Alvarez-Barrientos A, Arroyo J, Nombela C, Rivas C: **Identification of a nuclear export signal in the KSHV latent protein LANA2 mediating its export from the nucleus**. *Experimental cell research* 2005, **311**:96–105.

94. Nie Y: **Subcellular Distribution of ADAR1 Isoforms Is Synergistically Determined by Three Nuclear Discrimination Signals and a Regulatory Motif**. *Journal of Biological Chemistry* 2003, **279**:13249-13255.

95. Thyssen G, Li T, Lehmann L, Zhuo M, Sharma M, Sun Z: **LZTS2 Is a Novel  -Catenin-Interacting Protein and Regulates the Nuclear Export of  -Catenin**. *Molecular and Cellular Biology* 2006, **26**:8857-8867.

96. Kwon I, Lee J, Chang SH, Jung NC, Lee BJ, Son GH, Kim K, Lee KH: **BMAL1 Shuttling Controls Transactivation and Degradation of the CLOCK/BMAL1 Heterodimer**. *Molecular and Cellular Biology* 2006, **26**:7318-7330.

97. Ito S, Nagaoka H, Shinkura R, Begum N, Muramatsu M, Nakata M, Honjo T: **Activation-induced cytidine deaminase shuttles between nucleus and cytoplasm like apolipoprotein B mRNA editing catalytic polypeptide 1**. *Proceedings of the National Academy of Sciences of the United States of America* 2004, **101**:1975.

98. Itahana Y, Yeh ETH, Zhang Y: **Nucleocytoplasmic Shuttling Modulates Activity and Ubiquitination-Dependent Turnover of SUMO-Specific Protease 2**. *Molecular and Cellular Biology* 2006, **26**:4675-4689.

99. Kulisz A, Simon H: **An Evolutionarily Conserved Nuclear Export Signal Facilitates Cytoplasmic Localization of the Tbx5 Transcription Factor**. *Molecular and Cellular Biology* 2007, **28**:1553-1564.

100. Yi C, Wang H, Wei N, Deng XW: **An initial biochemical and cell biological characterization of the mammalian homologue of a central plant developmental switch, COP 1**. *BMC Cell Biology* 2002, **3**:30.

101. Bartholomeusz G, Wu Y, Seyed MA, Xia W, Kwong KY, Hortobagyi G, Hung MC: **Nuclear translocation of the pro-apoptotic Bcl-2 family member Bok induces apoptosis**. *Molecular carcinogenesis* 2006, **45**:73–83.

102. Yang Y: **Nucleocytoplasmic Shuttling of Receptor-interacting Protein 3 (RIP3): IDENTIFICATION OF NOVEL NUCLEAR EXPORT AND IMPORT SIGNALS IN RIP3**. *Journal of Biological Chemistry* 2004, **279**:38820-38829.

103. Thakurta AG: **Conserved Nuclear Export Sequences in Schizosaccharomyces pombe Mex67 and Human TAP Function in mRNA Export by Direct Nuclear Pore Interactions**. *Journal of Biological Chemistry* 2004, **279**:17434-17442.

104. Nishie T, Nagata K, Takeuchi K: **The C protein of wild-type measles virus has the ability to shuttle between the nucleus and the cytoplasm**. *Microbes and Infection* 2007, **9**:344–354.

105. Lim MA, Kikani CK, Wick MJ, Dong LQ: **Nuclear translocation of 3′-phosphoinositide-dependent protein kinase 1 (PDK-1): a potential regulatory mechanism for PDK-1 function**. *Proceedings of the National Academy of Sciences of the United States of America* 2003, **100**:14006.

# Supplementary materials

## Table S1: A list of 70 newly discovered NES-containing proteins

| UniProt_Id | Verified NES region | Reference |
|---|---|---|
| C8ZJE1 | 297-305; 20-28 | [31]; [32] |
| O00401 | 225-235 | [33] |
| O60729 | 393-402 | [34] |
| O70201 | 96-104; 89-98 | [35]; [36]; [37] |
| O92398 | 100-112 | [38] |
| O94916 | 1-19 | [39] |
| P02554 | 41-50 | [40] |
| P03409 | 189-202 | [41] |
| P05230 | 145-152 | [3] |
| P06748 | 94-102 | [42]; [43]; [44] |
| P08353 | 128-137 | [45] |
| P0CA03 | 139-148; 1-10 | [46] |
| P11142 | 394-401 | [47] |
| P11388 | 1017-1028; 1054-1066 | [48]; [49] |
| P15336 | 405-413 | [50] |
| P16110 | 241-256 | [51] |
| P16727 | 228-237; 359-366 | [52] |
| P17861 | 186-208 | [53] |
| P19508 | 41-60 | [54] |
| P20067 | 91-102 | [55]; [56] |
| P21952 | 367-376 | [57] |
| P22363 | 49-58 | [58] |
| P28289 | 127-136 | [59] |
| P30021 | 485-495 | [60] |
| P38483 | 173-196 | [61] |
| P38936 | 68-78; 102-119 | [62] |
| P42345 | 975-984; 1281-1289; 535-547 | [63] |
| P42858 | 2397-2406 | [64] |
| P50616 | 2-14 | [65] |
| P51587 | 1383-1393 | [66] |
| P53686 | 299-357 | [67] |
| P59636 | 46-54 | [68] |
| P67775 | 149-158 | [69] |
| P97318-2 | 152-160; 462-469 | [70] |

| | | |
|---|---|---|
| Q02085 | 132-143 | [71] |
| Q02880 | 1034-1044 | [49] |
| Q14145 | 272-312 | [72] |
| Q14494 | 251-260 | [73] |
| Q14678 | 613-622 | [74] |
| Q16236-2 | 537-546; 175-186 | [75]; [76] |
| Q18PE1 | 240-249 | [77] |
| Q5EBH1 | 372-379 | [78] |
| Q5GLC3 | 145-165 | [79] |
| Q60795 | 545-554 | [80] |
| Q61699 | 607-617 | [81] |
| Q68SB1-2 | 4-14 | [82] |
| Q6UB99 | 2415-2424 | [83] |
| Q75R42 | 189-198 | [84] |
| Q83414 | 82-91 | [85] |
| Q8IXJ6 | 41-51 | [86] |
| Q8K3Y6 | 284-291 | [87] |
| Q8N720 | 95-109 | [88] |
| Q95168 | 719-728; 728-738; 305-313 | [89] |
| Q96JZ2 | 199-211 | [90] |
| Q96T21 | 756-770; 634-657 | [91] |
| Q99152 | 37-46 | [92] |
| Q99AM3 | 551-560 | [93] |
| Q99MU3 | 128-137 | [94] |
| Q9BRK4 | 631-642 | [95] |
| Q9EPW1 | 142-152; 361-369 | [96] |
| Q9GZX7 | 190-198 | [97] |
| Q9HC62 | 317-332 | [98] |
| Q9PWE8 | 152-160 | [99] |
| Q9R1A8 | 237-247 | [100] |
| Q9UMX3 | 70-78 | [101] |
| Q9UNH5 | 355-364 | [34] |
| Q9Y572 | 255-264; 344-354 | [102] |
| Q9Y8G3 | 434-509 | [103] |
| Q9YZN9 | 76-85 | [104] |
| Q9Z2A0 | 382-391 | [105] |

**Table S2: The list of proteins annotated with dual localization ranked by the split NESsential trained by the combined dataset**

| UniProt_Id | Position | Probability |
|---|---|---|
| NAP1_YEAST | 334 | 0.692092 |
| PTP2_YEAST | 316 | 0.645572 |
| NOT2_YEAST | 127 | 0.604863 |
| R101_YEAST | 600 | 0.604778 |
| NOT3_YEAST | 146 | 0.572437 |
| RPB7_YEAST | 7 | 0.543616 |
| NOT1_YEAST | 665 | 0.5 |
| RAD5_YEAST | 338 | 0.478667 |
| LOS1_YEAST | 606 | 0.433603 |
| RPC3_YEAST | 641 | 0.415095 |
| SSB1_YEAST | 53 | 0.414102 |
| DBP2_YEAST | 321 | 0.407879 |
| PP12_YEAST | 54 | 0.407583 |
| NOT5_YEAST | 474 | 0.401357 |
| SKI3_YEAST | 300 | 0.393377 |
| ARG2_YEAST | 240 | 0.354822 |
| POP1_YEAST | 563 | 0.320881 |
| CTK3_YEAST | 197 | 0.319269 |
| DEP1_YEAST | 208 | 0.308044 |
| NB35_YEAST | 71 | 0.307779 |
| CC27_YEAST | 7 | 0.306051 |
| RMI1_YEAST | 57 | 0.303967 |
| RPB4_YEAST | 47 | 0.302519 |
| BMS1_YEAST | 134 | 0.279144 |
| ARP8_YEAST | 310 | 0.27765 |
| RPB6_YEAST | 148 | 0.27237 |
| FAP7_YEAST | 93 | 0.269712 |
| DCUP_YEAST | 244 | 0.256273 |
| DPB2_YEAST | 405 | 0.25325 |
| PR28_YEAST | 188 | 0.245672 |
| MSI1_YEAST | 119 | 0.243744 |
| C1TC_YEAST | 338 | 0.231486 |
| CB31_YEAST | 5 | 0.224899 |
| MTD1_YEAST | 310 | 0.220534 |
| ST20_YEAST | 88 | 0.219569 |
| LSM1_YEAST | 112 | 0.207213 |
| CEF1_YEAST | 428 | 0.188546 |
| SYYC_YEAST | 108 | 0.17009 |

| | | |
|---|---|---|
| LSM8_YEAST | 13 | 0.161685 |
| CTK1_YEAST | 165 | 0.160456 |
| NOT4_YEAST | 138 | 0.151759 |
| HAT2_YEAST | 55 | 0.145349 |
| UBC3_YEAST | 44 | 0.144261 |
| IST3_YEAST | 56 | 0.11431 |
| EGD1_YEAST | 51 | 0.109843 |
| RPD3_YEAST | 301 | 0.109094 |
| IPB2_YEAST | 45 | 0.108086 |
| UBC2_YEAST | 29 | 0.048252 |

**Table S3: The list of proteins annotated with nuclear localization ranked by the split NESsential trained by the combined dataset (Prob. ≥ 0.5)**

| UniProt_Id | Position | Probability |
|---|---|---|
| GLN3_YEAST | 723 | 0.925172 |
| HAP4_YEAST | 546 | 0.800667 |
| SPB1_YEAST | 665 | 0.783344 |
| NH10_YEAST | 45 | 0.764689 |
| ACE1_YEAST | 124 | 0.75475 |
| SMI1_YEAST | 500 | 0.712017 |
| HED1_YEAST | 54 | 0.70771 |
| RPC4_YEAST | 259 | 0.695314 |
| SNF4_YEAST | 314 | 0.693641 |
| RA18_YEAST | 207 | 0.688935 |
| ITC1_YEAST | 384 | 0.680225 |
| TAF7_YEAST | 578 | 0.677138 |
| SPT7_YEAST | 152 | 0.653795 |
| GAL4_YEAST | 64 | 0.639153 |
| MSN1_YEAST | 40 | 0.625432 |
| RRP1_YEAST | 118 | 0.620273 |
| PDR1_YEAST | 475 | 0.615326 |
| HAA1_YEAST | 536 | 0.61197 |
| MET4_YEAST | 604 | 0.607666 |
| ERB1_YEAST | 192 | 0.588147 |
| HIR1_YEAST | 419 | 0.586253 |
| CBF1_YEAST | 245 | 0.583998 |
| DRS1_YEAST | 649 | 0.568691 |
| NOP2_YEAST | 105 | 0.568534 |
| MAD3_YEAST | 381 | 0.564558 |
| ACE2_YEAST | 70 | 0.563049 |
| SIR4_YEAST | 979 | 0.562793 |
| SSF1_YEAST | 331 | 0.562084 |
| MCM3_YEAST | 834 | 0.55734 |
| PUT3_YEAST | 488 | 0.557014 |
| ADR1_YEAST | 766 | 0.555824 |
| MEI5_YEAST | 103 | 0.555519 |
| MP10_YEAST | 341 | 0.550755 |
| SWI1_YEAST | 951 | 0.546881 |
| ORC6_YEAST | 81 | 0.543549 |
| TAF9_YEAST | 35 | 0.542758 |
| PR40_YEAST | 344 | 0.537894 |
| MBP1_YEAST | 757 | 0.536536 |

| | | |
|---|---|---|
| TAFC_YEAST | 442 | 0.536245 |
| PR21_YEAST | 215 | 0.534033 |
| PFD6_YEAST | 15 | 0.526597 |
| T2FB_YEAST | 390 | 0.523427 |
| REF2_YEAST | 178 | 0.523374 |
| DPO2_YEAST | 64 | 0.52287 |
| MAD1_YEAST | 393 | 0.519251 |
| ARP5_YEAST | 563 | 0.51651 |
| UTP11_YEAST | 110 | 0.516043 |
| SKN7_YEAST | 492 | 0.512903 |
| UGA3_YEAST | 114 | 0.510215 |
| PR22_YEAST | 448 | 0.509755 |
| MKS1_YEAST | 370 | 0.508834 |
| MU81_YEAST | 145 | 0.505662 |
| MTR4_YEAST | 461 | 0.505412 |
| PIP2_YEAST | 610 | 0.5 |
| HIR2_YEAST | 481 | 0.5 |
| TF3A_YEAST | 337 | 0.5 |
| SEN2_YEAST | 215 | 0.5 |
| ORC1_YEAST | 208 | 0.5 |
| ASH1_YEAST | 166 | 0.5 |

**Table S4: The list of proteins annotated with cytosolic localization ranked by the split NESsential trained by the combined dataset (Prob. ≥ 0.5)**

| UniProt_Id | Position | Probability |
|---|---|---|
| SNF7_YEAST | 234 | 0.957903 |
| MDS3_YEAST | 1353 | 0.89723 |
| SYV_YEAST | 1053 | 0.662564 |
| NBP2_YEAST | 228 | 0.652946 |
| RNA1_YEAST | 347 | 0.634038 |
| BTN2_YEAST | 133 | 0.609133 |
| NTF2_YEAST | 54 | 0.578835 |
| STI1_YEAST | 191 | 0.569749 |
| KR11_YEAST | 379 | 0.567037 |
| CNS1_YEAST | 294 | 0.565512 |
| VPS3_YEAST | 521 | 0.541156 |
| CYPH_YEAST | 53 | 0.536443 |
| NMT_YEAST | 13 | 0.523817 |
| VPS5_YEAST | 510 | 0.5173 |
| YBP1_YEAST | 501 | 0.515131 |
| COAC_YEAST | 1190 | 0.510611 |
| GLRX_YEAST | 36 | 0.5 |

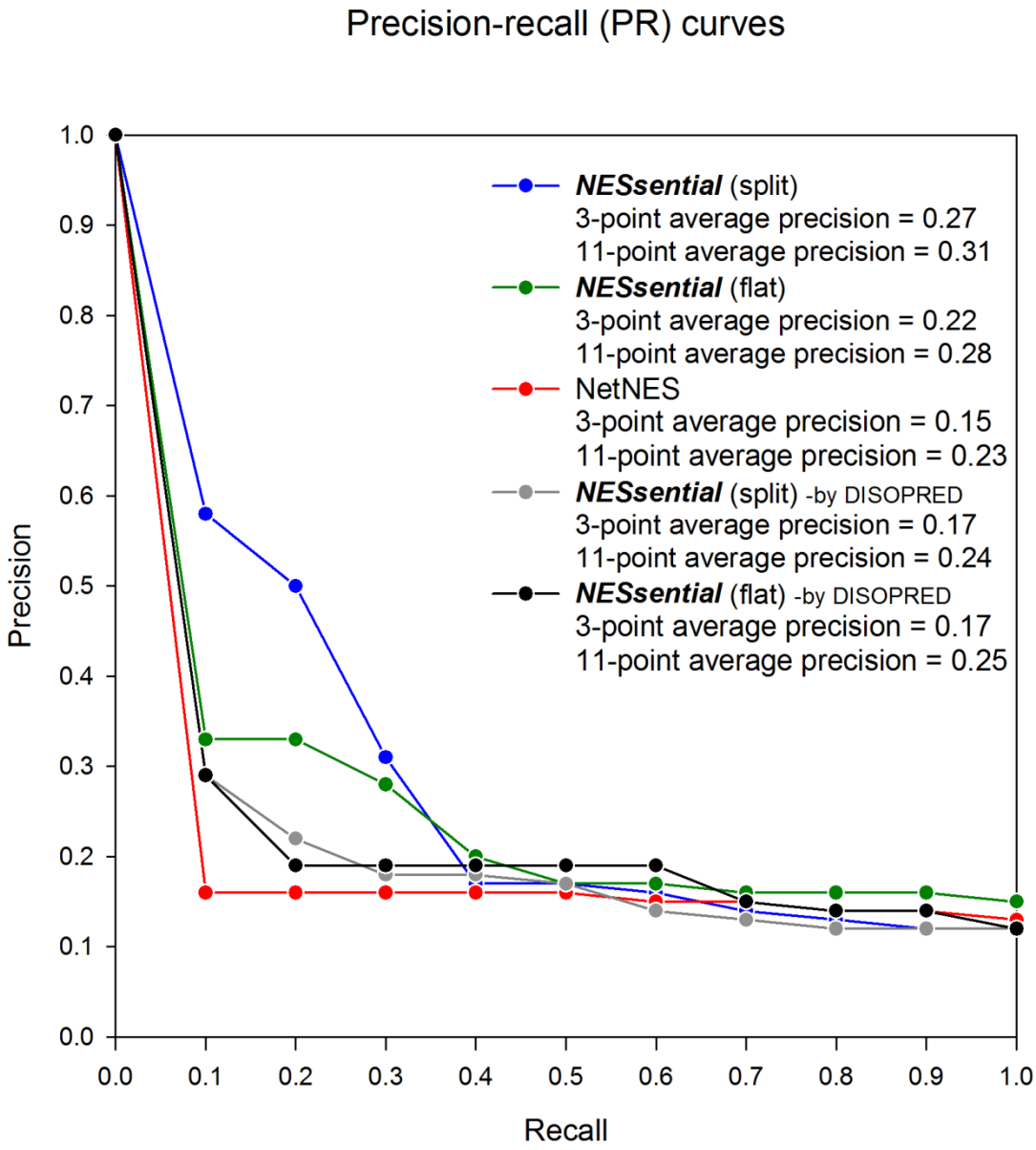



Figure S1: The precision-recall curves of different types of NESsential and NetNES. Black and grey solid curves represent for the performance of flat and split NESsential respectively while POODLE-L was replaced by another disorder predictor, DISOPRED.

## F-score and probability estimates

For the convenience of readers, we repeat the definition of the F-score given by Chen and Lin [http://www.csie.ntu.edu.tw/~cjlin/papers/features.pdf], and simply summarize how libsvm package extends SVM to give probability estimates given by Chang and Lin [http://www.csie.ntu.edu.tw/~htlin/paper/doc/plattprob.pdf] as follows:

**F-score:**

F-score is a simple measure of the discrimination of two sets of real numbers. Given training vectors $x_k$, $k = 1, \ldots, m$, if the number of positive and negative instances are $n_+$ and $n_-$, respectively, then the F-score of the $i$th feature is defined as:

$$F(i) \equiv \frac{\left(\bar{x}_i^{(+)} - \bar{x}_i\right)^2 + \left(\bar{x}_i^{(-)} - \bar{x}_i\right)^2}{\frac{1}{n_+ - 1}\sum_{k=1}^{n_+}\left(x_{k,i}^{(+)} - \bar{x}_i^{(+)}\right)^2 + \frac{1}{n_- - 1}\sum_{k=1}^{n_-}\left(x_{k,i}^{(-)} - \bar{x}_i^{(-)}\right)^2}$$

where $\bar{x}_i$, $\bar{x}_i^{(+)}$, $\bar{x}_i^{(-)}$ are the average of the $i$th feature of the whole, positive, and negative data sets, respectively; $x_{k,i}^{(+)}$ is the $i$th feature of the $k$th positive instance, and $x_{k,i}^{(-)}$ is the $i$th feature of the $k$th negative instance.

**Probability Estimates:**

The original SVM predicts only class label, NES motif or not in our case, without probability information. LIBSVM package extends SVM and supports a function to give probability estimates by a sigmoid function proposed by Platt et al. (2000):

$$\Pr(y = 1|x) \approx P_{A,B}(f) \equiv \frac{1}{1 + \exp(Af + B)}, \text{ where } f = \mathrm{f}(x). \qquad (1)$$

Let each $f_i$ be an estimate of $\mathrm{f}(x_i)$. The best parameter setting z* =(A*,B*) is determined by solving the following regularized maximum likelihood problem (with $N_+$ of the $y_i$'s positive, and $N_-$ negative):

$$\min_{z=(A,B)} \quad F(z) = -\sum_{i=1}^{l}\Big(t_i \log(p_i) + (1 - t_i)\log(1 - p_i)\Big), \qquad (2)$$

$$\text{for} \quad p_i = P_{A,B}(f_i), \text{ and } t_i = \begin{cases} \frac{N_+ + 1}{N_+ + 2} & \text{if } y_i = +1 \\ \frac{1}{N_- + 2} & \text{if } y_i = -1 \end{cases}, i = 1, \ldots, l.$$