Precise Nuclear Export Signal Detection in Protein Sequences Using Zero-Shot Segmentation and Multi-Modal Scoring

Noam Sabag Bickel and Shay Bergman

Abstract

Nuclear Export Signals (NES) are short leucine-rich motifs that direct protein export from the nucleus to cytoplasm via CRM1-mediated transport.
Traditional NES prediction methods rely on whole-sequence analysis, often lacking precision in localizing exact NES boundaries within longer proteins.
Recent advances in protein language models and change-point detection offer new opportunities for precise motif localization.
We developed a pipeline combining zero-shot protein segmentation using ESM2 based embeddings with multi-modal scoring that integrates consensus pattern matching, embedding similarity analysis and neural network classification.
Our method uses change-point analysis to identify domain boundaries (segments), followed by NES-specific scoring and optimal threshold optimization(for the multi-modal scoring).

The neural network component achieved the highest individual performance (97% AUC, 78.95% F1-score), It is very important to mention though, that the neural network classifier was evaluated on the training set, so its reported performance may be overestimated due to potential overfitting and should be interpreted with caution when considering generalization to new data. The consensus patterns contributed 78% AUC and embedding similarity 69% AUC. The integrated approach yielded 92% AUC with 95.3% accuracy and 81% precision at optimal thresholds, demonstrating effective complementarity between sequence-based patterns and learned representations. We think this method may perhaps enable precise NES localization. The integration framework can be adapted for other short functional sequences, advancing computational tools for protein functional annotation.

Introduction

Nuclear-cytoplasmic transport is a fundamental cellular process governing protein localization and function. Nuclear Export Signals (NES) are conserved sequence motifs that direct proteins from the nucleus to the cytoplasm through recognition by the CRM1 export receptor. These signals typically consist of leucine-rich sequences following specific consensus patterns established by Kosugi et al. and refined by subsequent structural studies.

Traditional computational approaches for NES prediction have focused on sequence-level classification, often analyzing entire proteins or fixed-length sliding windows. While these methods have contributed significantly to our understanding of nuclear export mechanisms, they face limitations in precisely localizing NES boundaries and distinguishing functional signals from false positives in complex protein sequences. The challenge is compounded by the variable nature of NES motifs and their embedding within diverse protein contexts.

Recent advances in protein language models, particularly the ESM and ProtT5 architectures, have demonstrated remarkable capabilities in capturing protein sequence relationships and functional patterns. Simultaneously, zero-shot segmentation techniques have emerged as powerful tools for identifying domain boundaries without requiring extensive training data. The integration of these approaches presents an opportunity to develop more precise and generalizable methods for motif detection.

Our primary aim was to develop a comprehensive pipeline for precise NES detection that combines zero-shot protein segmentation with multi-modal scoring(structural domain knowledge and neural net classification) to accurately identify NES boundaries within full-length protein sequences while minimizing false positive predictions.

Methods

We first started by establishing our segmentation method(based on zero-shot protein segmentation by Sangster et al.), we adjusted the algorithm to try to segment the proteins to segments that are around the size of known NESs(8-25 AA).
Using the segmentation, we adjusted the provided nesDB dataset on a segment basis(labeling the segments that contained NESs as 1, and those who didn't as 0) and based our work on this adjusted segmentation database.
Post segmentation, we used two main methods to score the segments(to then "pick" the segment or segments with the highest scores as the most likely to contain NESs).

The first was a structural, pattern analysis based approach:
We implemented comprehensive pattern matching based on established NES consensus sequences, incorporating:

- Kosugi Classification System: All six pattern classes (1a-1d, 2, 3) with position specific hydrophobic residue requirements.
- Reverse Pattern Detection: Bidirectional binding patterns identified by Lee et al., weighted at 50% of forward pattern scores.
- Structure-Based Patterns: PKI-class and Rev-class motifs derived from crystal structure analyses.
- Biochemical Features: Hydrophobic content optimization, proline penalties, and acidic flanking preferences.

Pattern matching employed sliding window analysis with weighted scoring: Kosugi Class 1a (4 points), Classes 1b-1d and 2 (3 points), Class 3 (2 points), reflecting experimental validation frequencies.

The second approach was to use a neural network classifier based on a regression approach, by trying to learn the underlying properties of the NES containing segments based on their ESM2 embeddings and then predicting the segments containing NES based on threshold calculated on a separate validation set.

The third approach was Embedding Similarity Analysis, Segment embeddings were compared to reference NES/non-NES peptide sets using log-fold difference scoring adapted from the preparation exercise. This approach provides discriminative power between NES and non-NES segments, with positive scores indicating greater similarity to validated NES peptides.

We then also combined both approaches to create a combined score: The three scores were normalized and combined to give each segment a final score, allowing us to rank the most likely NES candidates based on this combined approach.

Finally, we analyzed and evaluated the three methods(structural based, neural network and combined scores) using accuracy, precision and recall metrics combined with
F1, and ROC AUC.

Experiments and Visualization

Discussion

This project(somewhat)successfully demonstrated precise Nuclear Export Signal detection using zero-shot protein segmentation with multi-modal scoring.
As seen in the previous sections, we can see that we managed reasonable yet not perfect prediction of segments containing NESs with the best results coming by applying a neural net as our scoring function and then by combining it with the domain knowledge based approach(pattern matching) and embedding scoring.

The multi-modal approach provides methodological robustness through multiple evidence lines, interpretability via detailed pattern information and scalability for proteome-wide screening.
Although the Neural network evaluation on training data likely inflates performance estimates(as the database size is limited, using only test results would mostly likely provide less generalization), requiring cautious interpretation.
Consensus patterns remain biased toward well-characterized NES classes, potentially missing novel variants. Segmentation parameters represent sensitivity-efficiency trade-offs that may affect boundary detection.
We think that our biggest pitfall was in our segmentation methodology, as we couldn't tweak the segmentation to the size of NESs and it was "distracted" by other motifs(it could cut the NES into multiple segments).
So looking into future directions, we would have liked to focus more on the segmentation, to try and more accurately capture appropriate NES segments, we could also incorporate adaptive weight optimization by developing a neural network to automatically optimize combination weights based on segment characteristics, enabling context-dependent optimal combinations across diverse protein families.

In conclusion, our multi-modal NES detection approach represents a novel approach in computational motif discovery, combining classical pattern recognition with modern machine learning. We tried to provide a framework based on multiple complementary signals for NES identification with high precision through optimal threshold selection.
Future developments addressing validation limitations and implementing adaptive optimization will enhance reliability and broaden impact across computational biology applications.

References