Bioinformatics, 2019, 1–3
doi: 10.1093/bioinformatics/btz657
Advance Access Publication Date: 22 August 2019
Applications Note

Structural bioinformatics

pCRM1exportome: database of predicted CRM1-dependent Nuclear Export Signal (NES) motifs in cancer-related genes

Yoonji Lee[1], Jordan M. Baumhardt[2], Jimin Pei[3], Yuh Min Chook[2] and Nick V. Grishin[1,3,*]

[1]Department of Biophysics, [2]Department of Pharmacology and [3]Howard Hughes Medical Institute, University of Texas Southwestern Medical Center, Dallas, TX 75390, USA

*To whom correspondence should be addressed.

Associate Editor: Arne Elofsson

Received on June 5, 2019; revised on August 1, 2019; editorial decision on August 16, 2019; accepted on August 20, 2019

Abstract

Motivation: The consensus pattern of Nuclear Export Signal (NES) is a short sequence motif that is commonly identified in protein sequences, whether the motif acts as an NES (true positive) or not (false positive). Finding more plausible NES functioning regions among the vast array of consensus-matching segments would provide an interesting resource for further experimental validation. Better defined NES should also allow meaningful mapping of cancer-related mutation positions, leading to plausible explanations for the relationship between nuclear export and disease.

Results: Possible NES candidate regions are extracted from the cancer-related human reference proteome. Extracted NES are scored for reliability by combining sequence-based and structure-based approaches. The confidently identified NES candidate motifs were checked for overlap with cancer-related mutation positions annotated in the COSMIC database. Among the ~700 cancer-related sequences in the COSMIC Cancer Gene Census, 178 sequences are predicted to have possible NES motifs containing cancer-related mutations at their key positions. These lists are organized into our database (pCRM1exportome), and other protein sequences in the human reference proteome can also be retrieved by their UniProt IDs.

Availability and implementation: The database is freely available at http://prodata.swmed.edu/pCRM1exportome.

Contact: grishin@chop.swmed.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

CRM1-dependent nuclear export signals (NESs) direct the active transport between the nucleus and cytoplasm for many cellular proteins. Cancer cells can also utilize this pathway to avoid antineoplastic mechanisms. The intracellular nuclear export of either tumor suppressive proteins or drug targets can result in drug resistance due to overexpression of CRM1 (Turner *et al.*, 2012). Therefore, the prediction of the NES motifs, which consist of short peptide sequences with hydrophobic (Φ) and spacer (X) residues having specific patterns (Supplementary Fig. S1), is of great interest to aid in the discovery of anticancer agents. For the validated cargo proteins in two databases, NESdb (Xu *et al.*, 2012) and validNES (Fu *et al.*, 2013), we previously presented a combined sequence- and structure-based approach to analyze the motifs' accessibility, secondary structure and stability in the CRM1's NES-binding site (Lee *et al.*, 2019). Here, we apply this approach to cancer-related human proteome sequences and provide a comprehensive database of all NES consensus patterns. The NES motifs are plotted together with (i) the disorder propensity, (ii) known domain information, (iii) predicted secondary structures, (iv) binding scores of the segments in the CRM1's NES-binding groove and (v) cancer-related mutation positions. The pCRM1exportome database provides a list of sequence segments predicted to be highly plausible, cancer-related, NES motifs that would be valuable and interesting resource for the further experimental validation.

2 Materials and methods

2.1 Extraction of the NES consensus sequences

Homo sapiens reference protein sequences were retrieved from UniProt (Proteome ID: UP000005640). Mitochondrial proteins and extracellular proteins (which are soluble proteins with signal peptides) were excluded from the list by using UniProt sequence annotation in Feature Table. For NES consensus patterns, we utilized a modified version of the Kosugi consensus (Xu *et al.*, 2012, 2015) and also refer the empirical class priority (for example, class 1a is the most abundant class; see *Lee et al., 2019* for more details). If the Φ_2-Φ_4 positions of the extracted region overlap with experimental evidence in NESdb or validNES, it is considered to be an experimentally validated NES region.

2.2 Disorder propensity

The disorder propensity of the protein sequences is calculated using three different programs, DISOPRED3 (Jones and Cozzetto, 2015), SPOT-disorder (Hanson *et al.*, 2017) and IUPred2A (Meszaros *et al.*, 2018). Truly ordered and buried regions with high confidence are defined using the strict cutoff value [If a residue's disorder propensities predicted by both DISOPRED and SPOT-disorder are below 0.15, the residue is defined as ordered ('O'); if not, the residue is recorded as ('D')]. If the portion of 'D' mark is more than 90% for the segment and flanking regions, the location of the segment (loc_DISO) is defined as an ordered region ('ORD'). If 'O' is more than 90%, the location is determined as a disordered region ('DISO'). The other segments are considered as the ones located in the 'boundary' region.

2.3 Known domains

By using Batch CD-search tool (Marchler-Bauer and Bryant, 2004), the protein domain information was extracted. Four different databases, i.e. CDD (cdd v3.16), NCBI_Curated (cdd_ncbi v3.16), Pfam (oasis_pfam v3.16), SMART (oasis_smart v3.16), were searched with the expected value threshold of 0.01. The results were retrieved by the Concise mode. Transmembrane domains and coiled-coil regions are marked based on UniProt annotations.

2.4 Secondary structure elements

Secondary structure elements of the protein sequences are predicted by PSIPRED Version 4.02 (Jones, 1999). During the PSI-BLAST search (Altschul *et al.*, 1997) to find homologs, uniref90_2015_01 (Suzek *et al.*, 2015) database is used. The confidence level of the prediction is colored by a gradient from dark (high confidence) to light (low confidence).

2.5 Relative binding energy (E_{bind})

A given peptide sequence is fitted to the backbone coordinates of every template structure. By using Rosetta Backrub (Smith and Kortemme, 2008) and Relax (Conway *et al.*, 2014; Nivon *et al.*, 2013) modules, the complex, the protein itself and the free peptide structures are modeled separately with the same process. The binding energy (E_{bind}) is calculated by $E_{complex}$—$E_{protein}$—$E_{peptide}$. For calculating $E_{peptide}$, we utilized the lowest energy among the all different backbone fitted models. Among the various template-fitted models, the one with the lowest E_{bind} score is finally selected.

2.6 Cancer-related mutation mapping

COSMIC mutation and Cancer Gene Census data were retrieved from https://cancer.sanger.ac.uk/cosmic (release v89). The meaningful mutation positions were selected using the following criteria: (i) FATHMM prediction is 'pathogenic'; (ii) Mutation somatic status is 'Confirmed somatic variant'; (iii) Mutation description does not contain 'coding silent', 'frameshift' or 'nonsense'; and (iv) the given position is not annotated as SNP.

3 Database and results

3.1 Predicted CRM1-dependent NES motifs

For the human reference proteome sequences, we extracted all possible NES consensus patterns and analyzed their relationship with predicted protein ordered/disordered regions, known protein domains, predicted secondary structure elements (SSE) and cancer-related mutation positions. In the case of known cancer-related proteins, which are annotated in the Catalogue of Somatic Mutations in Cancer (COSMIC) Cancer Gene Census (CGC) project (Sondka *et al.*, 2018), we modeled the CRM1-NES peptide complex structures and calculated the stability (E_{bind}) of NES peptides at the CRM1's NES-binding groove (see *Lee et al., 2019* for the details and the validation of the resulting binding energies). The generated 3D models of CRM1-NES peptide complexes can be downloaded from our database. For the validation purpose, we also provide the calculation results of the human sequences with experimentally validated NES regions (see Supplementary Text and Supplementary Table S1 for details).

3.2 Mapping cancer-related mutation positions to candidate NES regions

To predict the relationship of NES with cancer, we mapped the cancer-related mutation positions using COSMIC resources (Tate *et al.*, 2019) to the extracted sequences. Among the vast amount of data, meaningful positions that affect the amino acid character were selected (see Section 2 for details). If the key hydrophobic (Φ) residues of the NES consensus pattern-matching segment are mapped to the COSMIC mutation position, the segment is flagged with 'cosmic_phi'. In cases where the Φ residue is mutated to a conservative Leu, Ile, Val, Met or Phe, the segment is annotated as 'cosmic_phi_to_LIVMF'. When the key hydrophobic residues are mutated to other amino acid types, the segments are considered to be possible cancer-related candidates. If the spacer residues are mapped to the mutation position, the segment was flagged with 'cosmic_spacer'. Since the presence of Pro in the C-terminal spacer ($\Phi_2XX\Phi_3X\Phi_4$) regions of NES was reported to abolish the NES function (Kosugi *et al.*, 2008), we marked this proline mutation at the spacer region as 'cosmic_spacer_Pro'.

3.3 Possible cancer-related proteins with CRM1-dependent NES motifs

By combining these sequence-based and structure-based approaches, each segment was scored via criteria of the STAR-system (Supplementary Fig. S2). The yellow stars represent the scores related to NES, and the green stars are for the Cancer-related scores. A consensus-matching sequence segment is scored for being a candidate NES under the following conditions: (i) it is not located in highly ordered regions (one yellow star if passed), (ii) it does not have a β strand in the middle of the sequence (one yellow star if passed), (iii) it has a good or medium E_{bind} score, i.e. $E_{bind} < -30.0$ (one yellow star if passed). A sequence segment is scored for being cancer-related as follows: (i) two green stars are assigned to the segment if the key positions in NES consensus overlap with the cancer-related mutation(s), i.e. mutation(s) of the key hydrophobic positions altering the amino acid types to other than Leu, Ile, Val, Met or Phe or mutation(s) of the C-terminal spacer residues to Pro and (ii) one green star is given to the segment where cancer-related mutations overlap with the segment, but do not occur in key positions, including non-Pro mutation(s) in C-terminal spacer residues. If all criteria are passed, then the segment is predicted to be cancer-related. Among the 679 cancer-related sequences (in COSMIC Cancer Gene Census), 178 sequences possess top-ranking NES motifs (3 yellow stars) with cancer-related mutations mapped to key positions (two green stars) (see Supplementary Text and Supplementary Tables S2 and S3). These entries are listed in the 'Cancer-NES' tab of the database. The list contains previously validated cancer-related proteins, such as MutLα (Brieger *et al.*, 2011), Smad4 (Watanabe *et al.*, 2000), FOXO (Calnan and Brunet, 2008) and Keap1 (Sun *et al.*, 2007). In each entry, the user can view the full plot of the disorder propensity, known domain information, secondary structure, binding energy scores (E_{bind}) and the mapped mutation positions (one example shown in Supplementary Fig. S3). The detailed information of the NES consensus pattern-matching segments is listed in the table. In the 'Human-NES' tab of the database, the plots of the human reference proteome other than CGC proteins can also be retrieved.

4 Conclusions

We have constructed a comprehensive database which contains possible NES motifs of the cancer-related sequences in human proteome. One can easily check the NES consensus-matching sequence segments in relation to protein domains, secondary structures, binding energies at the CRM1's NES-binding groove and cancer-related mutation positions.

Acknowledgements

The authors acknowledge the Texas Advanced Computing Center (TACC; http://www.tacc.utexas.edu) at The University of Texas at Austin for providing HPC resources. We thank Lisa N. Kinch and R. Dustin Schaeffer for critical reading of the manuscript.

Funding

This work was funded by the Cancer Prevention Research Institute of Texas (CPRIT) Grants RP170170 (N.V.G. and Y.M.C) and RP180410 (Y.M.C.), the National Institutes of Health Grant (GM127390 to N.V.G.) and Welch Foundation Grants (I-1532 to Y.M.C and I-1505 to N.V.G).

Conflict of Interest: none declared.

References

Altschul,S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Brieger,A. *et al.* (2011) A CRM1-dependent nuclear export pathway is involved in the regulation of MutLalpha subcellular localization. *Genes Chromosomes Cancer*, **50**, 59–70.

Calnan,D.R. and Brunet,A. (2008) The FoxO code. *Oncogene*, **27**, 2276–2288.

Conway,P. *et al.* (2014) Relaxation of backbone bond geometry improves protein energy landscape modeling. *Protein Sci.*, **23**, 47–55.

Fu,S.C. *et al.* (2013) ValidNESs: a database of validated leucine-rich nuclear export signals. *Nucleic Acids Res.*, **41**, D338–D343.

Hanson,J. *et al.* (2017) Improving protein disorder prediction by deep bidirectional long short-term memory recurrent neural networks. *Bioinformatics*, **33**, 685–692.

Jones,D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.*, **292**, 195–202.

Jones,D.T. and Cozzetto,D. (2015) DISOPRED3: precise disordered region predictions with annotated protein-binding activity. *Bioinformatics*, **31**, 857–863.

Kosugi,S. *et al.* (2008) Nuclear export signal consensus sequences defined using a localization-based yeast selection system. *Traffic*, **9**, 2053–2062.

Lee,Y. *et al.* (2019) Structural prerequisites for CRM1-dependent nuclear export signaling peptides: accessibility, adapting conformation, and the stability at the binding site. *Sci. Rep.*, **9**, 6627.

Marchler-Bauer,A. and Bryant,S.H. (2004) CD-Search: protein domain annotations on the fly. *Nucleic Acids Res.*, **32**, W327–W331.

Meszaros,B. *et al.* (2018) IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic Acids Res.*, **46**, W329–W337.

Nivon,L.G. *et al.* (2013) A pareto-optimal refinement method for protein design Scaffolds. *PLoS One*, **8**, e59004.

Smith,C.A. and Kortemme,T. (2008) Backrub-like backbone simulation recapitulates natural protein conformational variability and improves mutant side-chain prediction. *J. Mol. Biol.*, **380**, 742–756.

Sondka,Z. *et al.* (2018) The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nat. Rev. Cancer*, **18**, 696–705.

Sun,Z. *et al.* (2007) Keap1 controls postinduction repression of the Nrf2-mediated antioxidant response by escorting nuclear export of Nrf2. *Mol. Cell. Biol.*, **27**, 6334–6349.

Suzek,B.E. *et al.* (2015) UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, **31**, 926–932.

Tate,J.G. *et al.* (2019) COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res.*, **47**, D941–D947.

Turner,J.G. *et al.* (2012) Nuclear export of proteins and drug resistance in cancer. *Biochem. Pharmacol.*, **83**, 1021–1032.

Watanabe,M. *et al.* (2000) Regulation of intracellular dynamics of Smad4 by its leucine-rich nuclear export signal. *EMBO Rep.*, **1**, 176–182.

Xu,D.R. *et al.* (2012) NESdb: a database of NES-containing CRM1 cargoes. *Mol. Biol. Cell*, **23**, 3673–3676.

Xu,D.R. *et al.* (2015) LocNES: a computational tool for locating classical NESs in CRM1 cargo proteins. *Bioinformatics*, **31**, 1357–1365.