SCIENTIFIC REPORTS

OPEN

Structural prerequisites for CRM1-dependent nuclear export signaling peptides: accessibility, adapting conformation, and the stability at the binding site

Received: 13 February 2019
Accepted: 11 April 2019
Published online: 29 April 2019

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

Nuclear export signal (NES) motifs function as essential regulators of the subcellular location of proteins by interacting with the major nuclear exporter protein, CRM1. Prediction of NES is of great interest in many aspects of research including cancer, but currently available methods, which are mostly based on the sequence-based approaches, have been suffered from high false positive rates since the NES consensus patterns are quite commonly observed in protein sequences. Therefore, finding a feature that can distinguish real NES motifs from false positives is desired to improve the prediction power, but it is quite challenging when only using the sequence. Here, we provide a comprehensive table for the validated cargo proteins, containing the location of the NES consensus patterns with the disordered propensity plots, known protein domain information, and the predicted secondary structures. It could be useful for determining the most plausible NES region in the context of the whole protein sequence and suggests possibilities for some non-binders of the annotated regions. In addition, using the currently available crystal structures of CRM1 bound to various classes of NES peptides, we adopted, for the first time, the structure-based prediction of the NES motifs bound to the CRM1's binding groove. Combining sequence-based and structure-based predictions, we suggest a novel and more straightforward approach to identify CRM1-binding NES sequences by analysis of their structural prerequisites and energetic evaluation of the stability at the CRM1's binding site.

Active transport between the nucleus and cytoplasm is an essential regulatory mechanism for many cellular proteins. As a major nuclear exporter factor, chromosome maintenance protein 1 (CRM1; or exportin-1, XPO1) mediates nuclear export of hundreds of distinct cargo proteins by recognizing short sequence motifs called Nuclear Export Signal (NES)[1–3]. CRM1 shuttles between the nucleus and the cytoplasm, binds cargo molecules at high RanGTP levels inside the nucleus, traverses nuclear pore complex (NPC) as ternary cargo–CRM1–RanGTP complexes, and releases cargo into the cytoplasm upon hydrolysis of the Ran-bound GTP[4]. Since spatial re-localization of oncoproteins and tumor suppressor proteins is important in cancer cells, understanding of the NES can help the basic research about this process and can also help the discovery of anticancer agents[5].

Classical NES motifs in the early studies were referred to as a cluster of hydrophobic residues, mostly leucines (hence also called Leu-rich NES), within a 10–15 residue-long sequence motif[1,6,7]. Many years of research on various export cargoes and randomization-and-selection screens showed that more residue types, such as Ile, Val, Met, and Phe, are also allowed at the hydrophobic positions of the CRM1-dependent NES signals[8,9]. These hydrophobic residues (Φ) are spaced with various patterns following the consensus $\Phi 1\text{-}(x)_{2\text{–}3}\text{-}\Phi 2\text{-}(x)_{2\text{–}3}\text{-}\Phi 3\text{-}x\text{-}\Phi 4$, where x denotes any amino acid. Later, structural studies of the CRM1 bound to NES peptides revealed another hydrophobic pocket in CRM1 that can bind to one more hydrophobic amino acid (Φ0)[10,11]. This site is less restricted to hydrophobic residues compared to others. Until recently, the existing 11 consensus patterns were

[1]Department of Biophysics, University of Texas Southwestern Medical Center, Dallas, TX, 75390, USA. [2]Howard Hughes Medical Institute, University of Texas Southwestern Medical Center, Dallas, TX, 75390, USA. [3]Department of Pharmacology, University of Texas Southwestern Medical Center, Dallas, TX, 75390, USA. Correspondence and requests for materials should be addressed to N.V.G. (email: grishin@chop.swmed.edu)

defined by the peptide library-based study[9] and structural analyses of CRM1-NES complexes[11–14]. They consist of four to five hydrophobic residues (Φ0-Φ4; generally, L, I, V, M, and F) which are bound to the corresponding hydrophobic pockets (P0-P4) in CRM1. Based on the pattern of these Φ's and spacing sequences, the NES motifs are classified as class 1a, 1b, 1c, 1d, 2, 3, and 4. Additionally, compared to these classes, some peptides bind in the opposite (−) direction, making their Φ3-Φ4 positions bound to P0-P1 (class 1-reverse)[13]. Until recently, X-ray crystal structures of CRM1 bound to NES peptides of the 1a, 1b, 1c, 2, 3, 4, and 1a-reverse classes have been solved. Depending on the classes, the NES peptides showed distinct backbone conformations binding to the central portion of the hydrophobic groove of CRM1. One turn helix in the middle is remarkably conserved among all classes maintaining a hydrogen bonding with the Lys residue (Lys568) in human CRM1[14].

Modeling short motifs or patterns like NES is a major research area in bioinformatics. Since NES motifs are essential regulators of the subcellular location of proteins in relation to cancer, cell cycle, cell differentiation and other important aspects of molecular biology, prediction of the NES motif is of great interest but still remains a challenge. Until now, more than 300 experimentally identified protein cargoes are recorded in databases such as validNESs[15] and NESdb[16] and over 1000 putative CRM1 cargoes were identified in a recent proteomics study[17]. Based on the ever-growing repertoire of the protein cargoes of CRM1, many attempts were tried to employ machine learning approaches to decide whether a given sequence has a CRM1-dependent NES motif or not. Several computational tools, such as NetNES[8], NESsential[18], NESmapper[19], LocNES[20], Wregex[21], and NoLogo[22] have been developed to predict NES motifs. Most of them are sequence-based predictors based on consensus pattern matching and calculation of biophysical properties such as disordered propensity, secondary structure components, and solvent accessibilities. To capture the diversity of the NES sequences, the consensus patterns were generally applied in the form of regular expression or position-specific scoring matrix (PSSM). Unfortunately, NES patterns are quite commonly observed in a large portion of the proteome so that the prediction based on these consensus patterns results in a high false positive rate. Since a functional NES needs to be solvent-exposed and not buried in a globular fold, Kırlı *et al.* applied these criteria and pattern matching to identify NES motifs in a set of validated, new CRM1 cargoes and found that functional NES motifs still could not be identified in a significant portion of them[17]. Moreover, sequences of functional NES motifs appear to be more diverse than previously appreciated. A large portion of experimentally defined NES regions does not match the current consensus patterns[17]. As a solution to reduce the high false positive rate, other biophysical features such as disorder propensity, secondary structure component, and evolutionary conservation were incorporated into machine learning algorithms like support vector machines (SVM) or neural networks[8,20]. However, the false positive rates remain high. In addition to the ever-expanding NES patterns resulting in many false positives when used in NES prediction, the limited information about direct CRM1 binding of the annotated NES regions is detrimental to develop accurate predictors using available data sets. Therefore, predicting NES motifs using only protein sequence information seems to have limitations, and the combination with structure-based predictions could be a new strategy to distinguish NES motifs and false positives.

In this study, using validated cargo protein sequences in NESdb and validNES, we provide a comprehensive look-up table which contains the location of the NES consensus patterns with the disorder propensity plots, conserved domain information, and the predicted secondary structure. This information could be useful for determining the most plausible NES region in the context of the whole protein sequence and for suggesting possibilities for some non-binders of the annotated NES regions. In addition, for the first time, we adopted the structure-based prediction of the NES sequences bound to the CRM1's NES binding groove, using multiple crystal structures of CRM1-NES peptide as templates. For several experimentally validated NES peptides and false positive ones, we calculated the relative binding energy of the sequence segments at the CRM1's binding pocket, and the prediction reliability of these binding energies was validated by the experimental binding affinities. Combining sequence-based and structure-based predictions, we suggest the novel and more straight-forward approach to identify NES sequences that bind directly to CRM1.

Results and Discussion

Deducing NES consensus pattern-matching sequences in candidate cargo proteins. Using the validated cargo protein sequences in NESdb and validNES (which have Leptomycin B (LMB)-sensitive data as evidence of CRM1-dependency), we extracted the NES consensus pattern-matching sequence segments based on the modified version of the Kosugi consensus[16,20] as summarized in Fig. 1. All the possible consensus patterns are recorded and prioritized by the empirical class priority (see *Methods* for details). Based on these criteria, 4226 consensus-matching segments were extracted for 318 cargo protein sequences. Among them, 463 segments were treated as candidate NES motifs as they occur in regions that overlap to experimental evidence, and 3763 were treated as false positives (FPs). The experimental NES regions of 54 cargo proteins do not match the current consensus and are not considered in this study. Also excluded are four cargo proteins with no reported NES regions and five cargos with long reported NES regions (>25 residues) that do not have specific residues annotated. Among the consensus patterns, class 1a is the most abundant class (41%) as expected. Especially, compared to the false positive sequences, class 1a is observed more than twice as often in the candidate NES sequences. Classes 1c, 2, and 3 follow with 14~15%, class 1a-reverse is observed in 8.6%, and classes 1b, 1d, 4, or 1c-reverse seem to be quite rare (Fig. S1).

A comprehensive look-up table of NES patterns in NES cargo proteins. In order to make the NES motif to be accessible to CRM1-binding, the motif should not be located in the compactly folded protein domains. The NES motif may be located at the N-terminus, at the C-terminus, or within an unstructured region of an export cargo[11]. Therefore, for a precise prediction of the export signals, it is crucial to consider the motifs' location with respect to protein domains and disordered regions. For all possible NES consensus patterns of the cargo proteins that we extracted, we analyzed the relationship with the protein ordered/disordered regions,

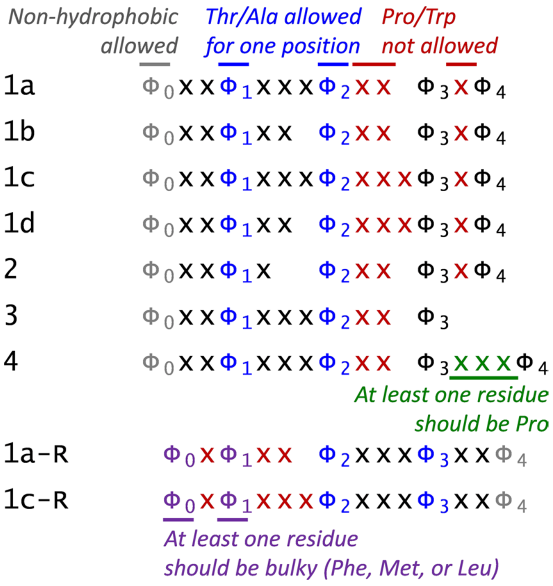

Non-hydrophobic allowed
Thr/Ala allowed for one position
Pro/Trp not allowed
1a Φ0 x x Φ1 x x x Φ2 x x Φ3 x Φ4
1b Φ0 x x Φ1 x x Φ2 x x Φ3 x Φ4
1c Φ0 x x Φ1 x x x Φ2 x x x Φ3 x Φ4
1d Φ0 x x Φ1 x x Φ2 x x x Φ3 x Φ4
2 Φ0 x x Φ1 x Φ2 x x Φ3 x Φ4
3 Φ0 x x Φ1 x x x Φ2 x x Φ3
4 Φ0 x x Φ1 x x x Φ2 x x Φ3 x x x Φ4
At least one residue should be Pro
1a-R Φ0 x Φ1 x x Φ2 x x x Φ3 x x Φ4
1c-R Φ0 x Φ1 x x x Φ2 x x x Φ3 x x Φ4
At least one residue should be bulky (Phe, Met, or Leu)

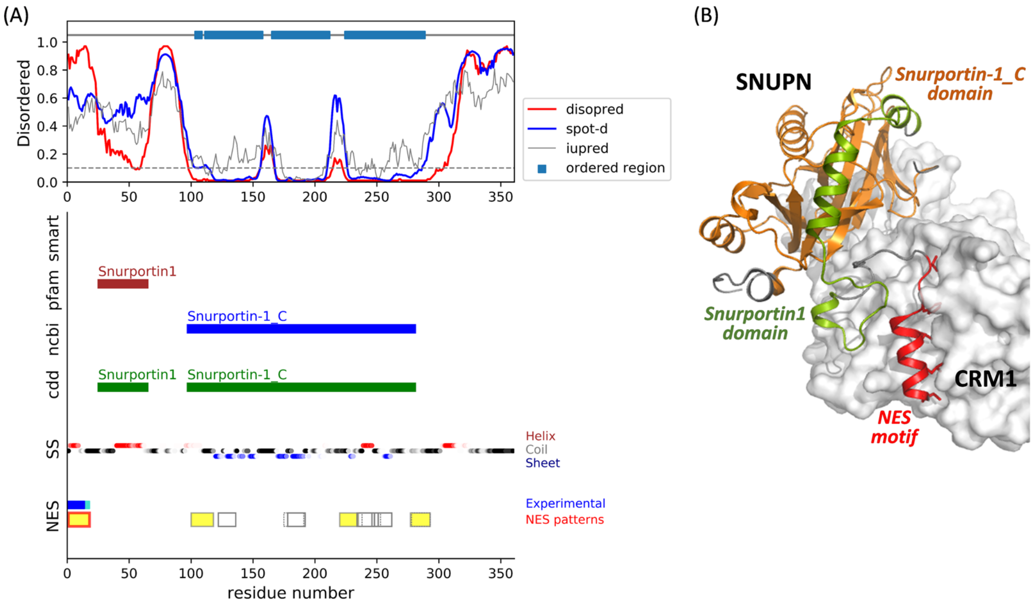
Figure 1. NES consensus patterns used in this study. For the hydrophobic positions, Φ_{1-4} are Leu, Ile, Val, Met, or Phe, and for the Φ_1 and Φ_2 positions, Thr or Ala is allowed for one position. Φ_0 is not restricted to the hydrophobic amino acids. In the reverse classes, the criteria are applied in the opposite direction, and one of the Φ_0 or Φ_1 should be Leu, Phe, or Met. The spacer residues (x) can be any amino acid, but several positions have exceptions. The spacers in Φ_2 $[X]_n\Phi_3X\Phi_4$ (or $\Phi_0X\Phi_1[X]_n\Phi_2$ in reverse classes) do not allow to have Pro or Trp. For class 4, at least one residue of the spacers in $\Phi_3XXX\Phi_4$ should be Pro to make a turn (as observed in the X-ray crystal structure of CRM1-X11L2 peptide).

known domains, and their predicted secondary structures, and provide a comprehensive online table. For a given full protein sequence, we plotted the disordered propensity, the location of the known domains, the predicted secondary structures, and all possible NES consensus regions (Fig. 2). For a given entry, the information annotated in NESdb or validNES, such as evidence of CRM1-dependency, mutation data, functional sequences or sites, is listed together. The locations of all NES consensus-matching segments are marked together with the experimentally validated regions (Fig. 2A, the bottom of the plot). The reference databases (NESdb, validNES, and UniProt), protein visualization tool (ProViz)[23] and the structure and model database (SWISS-MODEL repository)[24] are linked for user convenience, and the filter for easy look-up is also provided. This table could be useful for determining the most likely NES region in the context of a whole protein sequence. The online table is accessible via: http://prodata.swmed.edu/nes_pattern_location/.

NES candidates in the disordered or ordered regions. Even if a sequence motif can be fitted to the NES consensus, a motif that is located deep in the globular fold can hardly bind to CRM1 unless the region unfolds. In some cases, it may be possible to unfold and bind, but we assume that these cases would be very limited. Also, short linear interaction motifs like NES motifs have been proposed to be locally disordered to facilitate dynamic interactions with their binding partners, and the NES prediction algorithms have used disorder context to help distinguish correct NES motifs from false predictions[18,20]. However, NES motifs do not necessarily have to locate in the fully disordered region. Indeed, we have observed that some NES candidates are located in the fully disordered regions, but others are located next to ordered or "boundary" regions. Therefore, we employed the disorder propensity as a pre-filter to remove the segments located in the "highly" ordered regions.

Various computational tools have been developed for analyzing potential intrinsic disorder of protein sequences and were quite successful owing to clear association between disordered propensity and sequence features such as low complexity or high aromatic composition. We utilized DISOPRED3[25] and SPOT-disorder[26], which use homologous sequences' alignment-based profiles for detecting disordered regions, and IUPred2A[27] which is much faster since it does not rely on the sequence alignment. Disordered regions for some proteins are quite differently predicted depending on the programs. In order to define ordered and buried regions with high confidence, we applied strict cutoff values (~0.1) to decide the order/disorder border lines (note that the most of the programs' cutoff value for disordered regions are ~0.5). If a residue's disorder propensities predicted by both DISOPRED3 and SPOT-Disorder are below 0.1, the residue is defined as in highly ordered region (note that the predicted values by IUPred2A are also recorded for the reference).

As shown in Fig. 3A, 55% of the NES candidate motifs are located in the disordered region, and 37% are found in the boundary region between the ordered and disordered parts. Only 8% of the NES candidate motifs are located in the highly ordered region. Among the 361 candidate motifs, 37 segments (for 20 cargo proteins) are located in the highly ordered region which may have less possibility to be accessible to CRM1 binding. For example, HDAC1 (uniport ID: Q13547) has a reported NES motif with a mutation data (L158A/L161A/L164A) for nuclear export[28]. This region can be fitted to the classes 1c, 2, or 3, but it is located in the highly ordered region.



(C)

#	candidates	start#	sequence	secondary	class	multi-pattern	loc_DISO	loc_CDD	beta
1	cand_D	1	MEELSQALASSFSVSQ +***+++*++*++++	HHHHHHHHCCCCCCCC	c1c-5-Nt	uniq	DISO		0.0
2	fp_D	100	NQLMLSEWLIDVPSDLGQ	HHHHHHHHCCCCCCCCCC	c4-4	uniq	boundary	boundary|Snurportin-1_C	0.0
3	fp_beta_D	122	VVCPVGKRALIVAS	EEEECCCEEEEEEE	c3-AT	uniq	boundary	MID|Snurportin-1_C	0.571
4	fp_beta_D	175	IYNEVNQTYYVLDVMC	EEECCCCEEEEEEEEE	c1d-AT-4	multi	boundary	MID|Snurportin-1_C	0.571
5	fp_beta_D	178	EVNQTYYVLDVMCW	CCCCEEEEEEEEEE	c3-AT	multi-selected	boundary	MID|Snurportin-1_C	0.857
6	fp_D	220	EKTKLNPFKFVGLKN	CCCCCCCCEEEECCC	c1b-4	uniq	boundary	MID|Snurportin-1_C	0.429
7	fp_D	234	NFPCTPESLCDVLS	CCCCCHHHHHHHHC	c3-AT	multi	boundary	MID|Snurportin-1_C	0.0
8	fp_D	234	NFPCTPESLCDVLSMDF	CCCCCHHHHHHHHCCCC	c1c-AT-5	multi-selected	boundary	MID|Snurportin-1_C	0.0
9	fp_D	238	TPESLCDVLSMDFPF	CHHHHHHHHCCCCCC	c1b-4	multi	boundary	MID|Snurportin-1_C	0.0
10	fp_O	246	LSMDFPFEVDGLLFYH	HCCCCCCCCCEEEEEE	c1a-4	multi-selected	ORD	boundary|Snurportin-1_C	0.143
11	fp_O	248	MDFPFEVDGLLFYH	CCCCCCCCEEEEEE	c2-4	multi-selected	ORD	boundary|Snurportin-1_C	0.286
12	fp_D	277	LRPYMVSDVLGVAVPA	CCCCCCCCCCCCCCCC	c1a-4	multi-selected	boundary	boundary|Snurportin-1_C	0.0
13	fp_D	278	RPYMVSDVLGVAVPA	CCCCCCCCCCCCCCC	c1b-4	multi	boundary	boundary|Snurportin-1_C	0.0

Figure 2. Location of the NES consensus patterns in Snurportin-1. (**A**) Disordered propensity, conserved domain information, predicted secondary structure, and the location of the consensus patterns are plotted together. The defined ordered region (by the cutoff value of 0.1; gray dashed line) is represented by the sky-blue box at the top. The regions of the conserved domains annotated in smart, Pfam, NCBI-curated, and CDD are marked in the middle. The predicted secondary structures (SS) were colored by red, black, and blue for α-helix, coil, and β-strand, respectively. The gradient of the color corresponds to the confidence level of the prediction. For the NES regions, experimentally validated regions are displayed in blue (with mutation data annotated in NESdb) and cyan (annotated as a functional sequence in NESdb or as a site in validNES). All the consensus pattern matching segments are located at the bottom. Segments not in the ordered regions and without β-strand predictions in the middle are highlighted in yellow. The red boxes are the pattern-matching segments overlapping with experimental evidence. (**B**) The crystal structure of CRM1-SNUPN complex structure (PDB id: 3GB8)[12]. SNUPN is displayed by the cartoon, and the validated NES motif, Snurportin1 domain, and Snurportin-1_C domain are colored in red, green, and orange, respectively. CRM1 is represented by a white surface. (**C**) The list of the pattern-matching sequences in SNUPN. In the 'candidates' column, NES candidates and false positives are annotated with "cand" and "fp," respectively. If the segment is located in the disordered or boundary region, it is flagged with "_D" while in the ordered region, it is flagged with "_O." If the segment's β-strand content is over 0.5, it is flagged with "_beta." In the 'sequence' column, hydrophobic positions are colored in red, and the positions with the experimental evidence are marked with '*' (mutation) and '+' (functional sequence in NESdb or sites in validNES). The values in 'diso,' 'spotd,' and 'iup' are the average disordered propensity for the segment calculated by DISOPRED3, SPOT-Disorder, and IUPRED2A, respectively. The locations with respect to disordered/ordered region or conserved domains are listed in the 'loc_DISO' and 'loc_CDD' columns. 'beta' is for the β-strand content in the middle of the segment.

The crystal structure of HDAC1 (PDB ID: 4bkx) showed that this segment is buried in the globular domain and seems unlikely to be accessed by CRM1 (Fig. 4A). Note that in case of its homolog HDAC5, the candidate NES motif ($_{1081}$EEAE**T**VSA**M**AL**L**SVGA$_{1096}$, class 1a) is located in the disordered region after the conserved Hist_deacetyl domain and found to directly bind to CRM1. The similar region (after the Hist_deacetyl domain) in HDAC1 ($_{358}$**Y**LEK**I**KQR**L**FEN**L**R**M**LP$_{374}$, class 1c) could be also considered as a possible NES motif of HDAC1. Table S1 lists the NES candidate motifs located in the highly ordered region and Fig. 4A,C shows some examples for these segments in the available 3D structures.

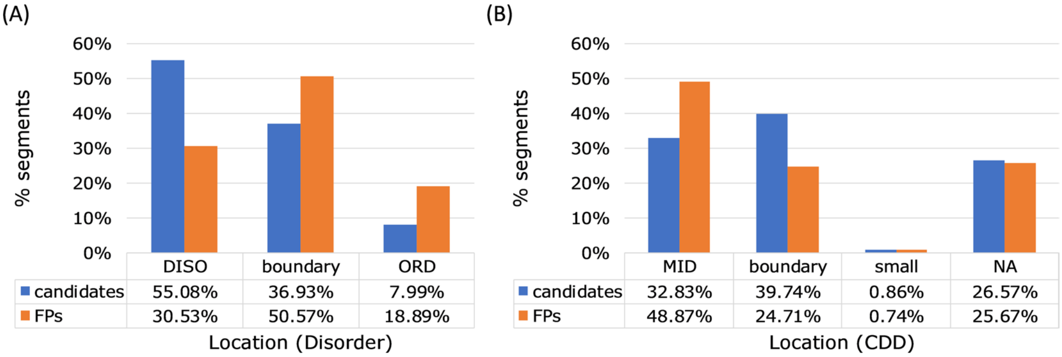


Figure 3. Location of the candidate NES and false positive sequences. (**A**) Location with respect to the disordered or ordered regions. DISO: located in the disordered region; boundary: located at the end of the highly ordered region; ORD: located in the highly ordered region. (**B**) Location with respect to the known domains annotated in CDD. MID: located in the middle of the domain; boundary: located at the end of the domain; small: located in the small domain (<50 residues); NA: located in the region with no annotated information.

In case of the false positives, the segments located in the highly ordered region is 19%, a larger percentage than those of the candidate NES motifs (note that the segments in the ordered region are far lower than those in the disordered region since we use the stringent cutoff for defining ordered region). The false positives in the disordered or boundary regions are 31% and 51%, respectively.

CDD domains and NES locations. To analyze the candidate NES motifs' location with respect to the conserved regions, we extracted the conserved domain information for the cargo protein sequences using the four different databases, i.e., SMART, Pfam, NCBI-curated, and Conserved Domain Database (CDD). As shown in Fig. 3B, only 33% of the candidate NES regions are located in the middle of the CDD domains, and 40% is in the boundary region. It seems that the NES regions do not necessarily locate in the protein domains. Rather, the known domains are often considered to form folding units, masking the possible motifs from binding other proteins. In case of the false positives, more than half are located in the middle of the known domains. It may be because the hydrophobic residues are commonly located in the protein core or domains.

Secondary structure components of the NES peptides. Crystal structures of CRM1-bound NES peptides have been resolved for the classes 1a, 1a-reverse, 1b, 1c, 2, 3, and 4. They showed distinct backbone conformations that match their hydrophobic positions to the corresponding hydrophobic pockets in CRM1. Structural analysis, as well as secondary structure prediction of NES motifs, suggest that most NES motifs contain α-helices or helix-to-extended conformation[12–14]. The class 1d is also expected to have helix-strand, and other reverse (−) classes are likely the reverse of their (+) counterparts[14]. The common feature of the backbone conformations among the classes is one turn of helix at the region from Φ2 to Φ3[14].

In our analysis of the 361 candidate motifs, 36 segments (for 23 cargoes) have a β-strand conformation in the middle (β-strand contents of the middle part is >50%) (Table S2). Among them, 11 segments were confirmed to have β-strands in the available X-ray or solution structures. For example, NPM has two reported NES regions, but both of them are predicted to form β-strands in the middle of the segments. As shown in Fig. 4B, the two segments are both β-strands located in the middle of the jelly-roll fold. Indeed, both regions were also reported to be quite weak binders of CRM1[29] and the sequence of 42–61 failed to bind CRM1 in GST-pulldown assay (Chook Lab, unpublished results; annotated in NESdb). The candidate NES region in TDP-43 is also located in β-strands within a folded globular RRM domain, and it is recently validated to be a non-binder to CRM1 rather it is exported by passive diffusion[30]. For six segments, there is no experimentally determined structure, but homology models showed the β-strands for the segments. For 17 segments, no structural information is available. For two segments, the conformation in the modeled structures (with sequence identities of 79% and 98%, respectively) are found to be helix reflecting the limitation of the secondary structure prediction.

Evaluation of the stability of the NES peptides at the CRM1 binding groove based on structure modeling. Recent structural works of CRM1 complexed with various cargo sequences expand the possible consensus patterns[13,14]. Also, the NES-binding site in RanGTP-bound CRM1 is found to be quite rigid, and the peptides display CRM1-dependent NES activity only if their backbone conformations can place a sufficient number of the hydrophobic residues into the CRM1's binding groove[11]. The adapting conformation of the peptides can be efficiently analyzed by structure-based modeling methods so that the application of the structural information can advance more accurate NES prediction.

Using the reported NES peptides with experimental binding affinities[14,31] as a benchmarking set (Table 1), we evaluated the binding energy (E_{bind}) for a given peptide sequence at the CRM1 groove (see *Methods* for details). Binding energy can be assumed as relative stability of the protein(CRM1)-peptide(NES) complex structure compared to the protein itself and free peptide. The lower the binding energy, the higher the possibility for the peptide segments to bind at CRM1. Multiple crystal structures of CRM1-NES peptide (super PKI and MVM-NS2 for

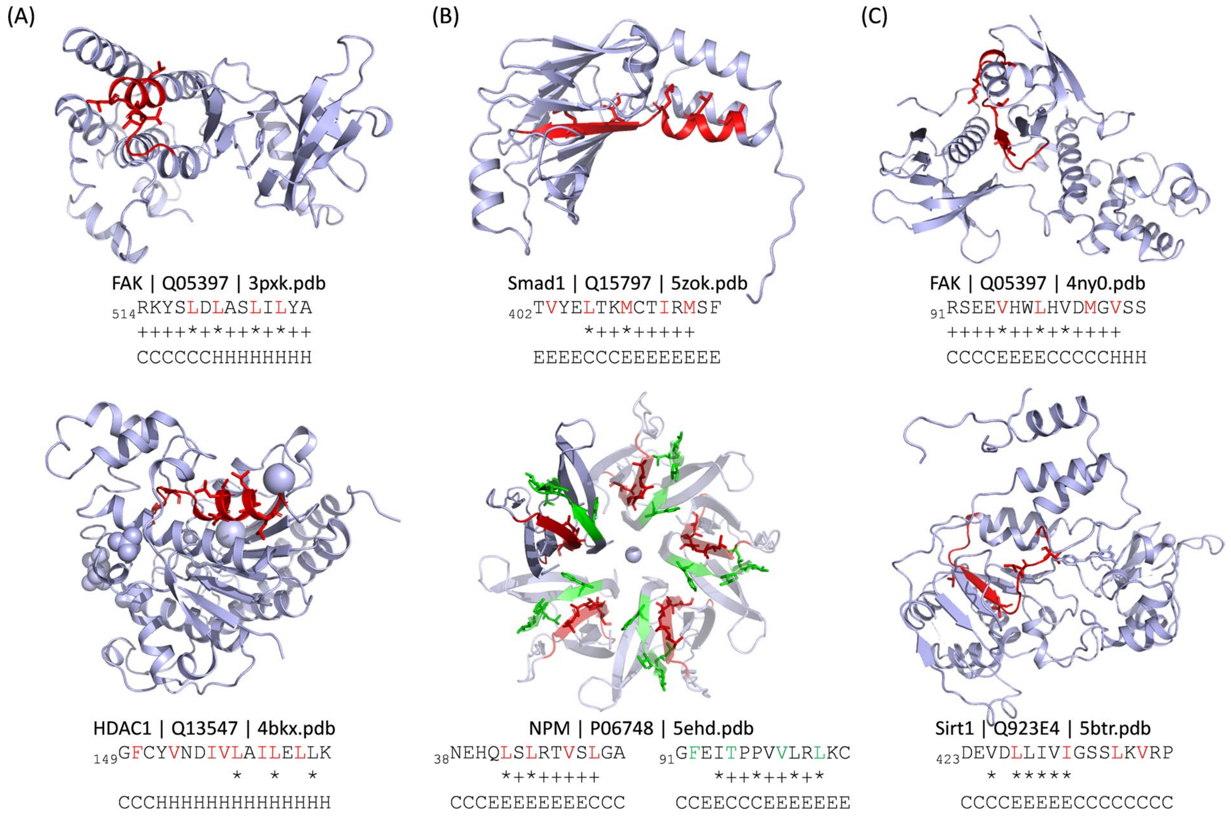


Figure 4. Examples of possible non-binders to CRM1. (**A**) Segments located in the ordered globular fold. (**B**) Segments with β-strands in the middle. (**C**) Segments located in the ordered region and have β-strands in the middle. The hydrophobic residues are colored in red or green in the sequences and displayed as sticks in the structures.

classes 1a; FMRP-1b for class 1b; SNUPN for class 1c; FMRP and SMAD4 for class 2; HIV-Rev for class2-rev type; X11L2 for class 4; and CPEB4 for class 1a-reverse; class 1a templates can be used to fit class 3 NES peptides) were utilized as templates. The model generation and energy calculation process are summarized in Fig. 5A.

Final model structures showed that all classes were predicted well with their Φ residues bound to the corresponding hydrophobic pockets (Fig. 5B). The calculated E_{bind} selected the right template for each class, and it can be utilized to find the most plausible class when multiple consensus patterns are found in one segment. The calculated E_{bind} values correlated quite well to the experimental K_D values (Fig. 6, left; R^2~0.63; *Pearson's r*~0.79 with $p = 2e - 6$). However, in the case of the two PKI mutant peptides which have extremely low binding affinities, the E_{bind} scores are not quite distinguishable from those of the weak binders such as SNUPN, SMAD4, and HPV-E7. In case of the PKI double mutant peptides, we found a large interface cavity at the binding interface with CRM1 (Fig. S3A), but this feature, definitely detrimental to binding, is not well reflected in the modeling process or energy calculation. To penalize the interface cavity of the complex structure, residue solvent accessibility (RSA) for key interface residues (Fig. S3B) is calculated using the NACCESS program[32] and treated as another scoring term. The RSA-corrected E_{bind} scores (E_{bind}^{RSA}) is obtained by calculating $E_{bind}^{RSA} = E_{bind} + w{\cdot}RSA$ (w is the weight for the RSA term and is optimized to maximize the correlation) (Fig. 6, middle). E_{bind}^{RSA} gave improved correlation (Fig. 6, right; R^2~0.73; *Pearson's r*~0.86 with $p = 5e\text{-}8$).

For comparison, several false positive sequences that can be fitted to NES consensus but are experimentally validated as non-binders (determined by pull-down binding assay)[13,33] are subjected to modeling with the same procedure. Interestingly, these false positives showed significantly higher E_{bind} scores reflecting their low binding affinities at the CRM1 binding groove. Notably, the peptides such as COMMD1 ($_{164}$DE**V**K**V**NQ**I**LKT**L**SE**V**EES$_{181}$) and ELF3 ($_{111}$R**L**V**F**GP**L**GDQ**L**HAQLR$_{126}$) were not fitted to the right template (i.e., the lowest E_{bind} complex is not the class 1a-R structure). It suggests that these sequences could be energetically unstable when their backbone conformations are fitted their hydrophobic residues to CRM1 hydrophobic pockets. In case of the false positive peptides fitted to the right template (Fig. 7), the backbone conformation and the Φ residues may appear to be pretty similar to the true positive ones; however, they showed inferior binding energies. In some cases, such as Cyclin D1 (Fig. 7A, middle) or FGF1 (Fig. 7C, right), the backbone conformation seems to be not maintained well when presenting the Φ side chains into the pockets.

We expect the merit of this structure-based, energy-based method is to discriminate true positive and false positive with similar sequence patterns, by analyzing energetic differences at the CRM1 binding site via full-atom modeling. This atomic-level energetic analysis cannot be deduced by using the only sequence. In this perspective, our method would suggest novel approaches to find the CRM1-binding NES motifs. We cannot ignore the fact

Protein	Class	NES sequence	K_D (nM)§	ref.
MVM NS2	1a	$_{77}$ST**V**DE**M**TKK**F**GT**L**T**I**HD$_{93}$	2	31
*super PKI	1a	$_{34}$N**L**NE**L**ALK**L**AG**L**D**I**NK$_{49}$	4	31
PKI	1a	$_{34}$NSNE**L**ALK**L**AG**L**D**I**NK$_{49}$	34	31
ADAR1	1a	$_{121}$RG**V**DC**L**SSH**F**QE**L**S**I**YQ$_{137}$	69	31
MEK1	1a	$_{28}$TN**L**EA**L**QKK**L**EE**L**E**L**DE$_{44}$	70	31
Pax	1a	$_{264}$RE**L**DE**L**MAS**L**SD**F**K**F**MA$_{280}$	700	31
*CPEB4-R	1a	$_{395}$RM**I**DI**L**SSE**L**SH**M**D**F**TR$_{379}$	710	31
NPMmutA	1a	$_{278}$MTDQEA**I**QD**L**CLA**V**EE**V**S**L**RK$_{298}$	790	31
HDAC5	1a	$_{1081}$EAE**T**VSA**M**AL**L**S**V**G$_{1095}$	1600	31
p73	1a	$_{364}$NFEI**L**MK**L**KES**L**EL**M**E**L**VP$_{382}$	2000	31
*hRio2-R	1a	$_{405}$GK**I**EE**L**AQN**F**ET**M**E**F**SR$_{389}$	2600	31
Stradα	1a	$_{413}$G**I**FG**L**VTN**L**EE**L**E**V**D$_{427}$	10300	31
*FMRP-1b	1b	YLKE**V**DQ**L**RA**L**ER**L**Q**I**D	3000	14
SNUPN	1c	$_{1}$**M**EE**L**SQA**L**ASS**F**S**V**SQDLNS$_{20}$	12500	31
HPV E7	1c	$_{73}$HVD**I**RT**L**EDL**L**MGT**L**G**I**VC$_{91}$	34000	31
HIV Rev	2	$_{73}$LQ**LP**P**L**ER**L**T**L**DC$_{85}$	1180	31
FMRP	2	$_{424}$LKE**V**DQ**L**R**L**ER**L**Q**I**D$_{438}$	2000	31
SMAD4	2	$_{134}$ERVVSPG**I**D**L**SG**L**T**L**Q$_{149}$	4600	31
mDia2	3	$_{1157}$SVPE**V**EA**L**LAR**L**RA**L**$_{1171}$	1600	31
CDC7	3	$_{456}$QD**L**RK**L**CER**L**RG**M**DSSTP$_{473}$	20000	31
X11L2	4	$_{55}$SS**L**QE**L**VQQ**F**EA**L**PGD**L**V$_{72}$	1500	31
CPEB4	1a-R	$_{379}$RT**F**D**M**HS**L**ESS**L**ID**I**MR$_{395}$	800	31
hRio2	1a-R	$_{389}$RS**F**E**M**TE**F**NQA**L**EE**I**KG$_{405}$	2800	31
*PKImut1 (I47A)	—	$_{34}$NSNE**L**ALK**L**AG**L**DANK$_{49}$	150000	31
*PKImut2 (L42A/L45A)	—	$_{34}$NSNE**L**ALKAAGAD**I**NK$_{49}$	900000	31
†APC	1a	$_{163}$AQ**L**QN**L**TKR**I**DS**L**P**L**$_{174}$	(—)	33
‡Cyclin D1	1a	$_{281}$VD**L**AC**T**PTD**V**RD**V**D**I**$_{295}$	(—)	—
APRIL	1b	$_{106}$LEP**L**KK**L**EC**L**KS**L**D**L**$_{120}$	(—)	33
‡hTERT	1c	$_{965}$KAGRN**M**RRK**L**FGV**L**R**L**KC$_{982}$	(—)	—
DcpS	1c	$_{136}$TEKH**L**QKY**L**RQD**L**R**L**$_{150}$	(—)	33
Cdk5	2	$_{133}$LINRNGE**L**K**L**AD**F**G**L**$_{147}$	(—)	33
†FGF1	2	$_{138}$THYGQKA**I**L**F**LP**L**P**V**$_{152}$	(—)	33
COMMD1	3	$_{171}$ILKT**L**SE**V**EES**I**ST**L**$_{185}$	(—)	20
DEAF1	1a-R	$_{452}$SW**L**Y**L**EE**M**VNS**L**LNTAQQ$_{469}$	(—)	13
SGN5	1a-R	$_{221}$YA**L**E**V**SY**F**KSS**L**DRKLL$_{238}$	(—)	13
†COMMD1–2	1a-R	$_{164}$DE**V**K**V**NQ**I**LKT**L**SE**V**EES$_{181}$	(—)	13
†ELF3	1a-R	$_{111}$R**L**V**F**GP**L**GDQ**L**HAQLR$_{126}$	(—)	13

Table 1. Peptide sequences of the validated NES motifs or false positives used in the structure-based modeling. *Engineered or mutated (underscored residues are the ones inserted or mutated). †Do not fit the consensus in Fig. 1 (due to Pro or do not have a bulky residue at Φ3/4 in the class 1a-R). ‡Unpublished data. §(—) means no binding determined by pull-down binding assay.

that the interaction between CRM1 and a whole cargo protein can be more than that of the CRM1-NES peptide[10]; however, it is extremely difficult to consider extra contacts between CRM1 and cargo's whole structure which may be different depending on each cargo. Based on our previous result describing the strength of the CRM1-NES peptide interaction correlated to the nuclear export activity[31], we assume that the energy prediction between CRM1 and NES peptide is a practical strategy.

For evaluating the performance, we compared our results to those of other sequence-based methods, i.e., NetNES[8], NESmapper[19], and LocNES[20] (Figs S4–S20). Using the whole sequences of 17 proteins in Table 1, we extracted 19 positive cases (regions annotated as NES motifs in the NESdb or validNES database with mutational evidence) and 341 negative cases (non-NES regions with consensus pattern-matching). As shown in Table S3, E_{bind} score performs the same as LocNES in terms of recall rate (both predicts 17 true positives out of 19 experimentally verified NES cases). On the other hand, E_{bind} outperforms LocNES in terms of specificity and false positive rate. E_{bind} recorded 23 cases of false positives while LocNES predicted nearly the double amount of false positives (40 cases). NetNES showed better specificity (true negative rate (TNR): 0.988) than our method (TNR: 0.933). However, its recall rate (sensitivity or true positive rate (TPR): 0.474) was much lower than our method (TPR: 0.895). Our method seems to work well enough compared to these available methods. It effectively decreases false positives while maintaining a high recall rate, showing the best performance with respect to the balance of precision & recall (F_1 score), and effectiveness (DOR).

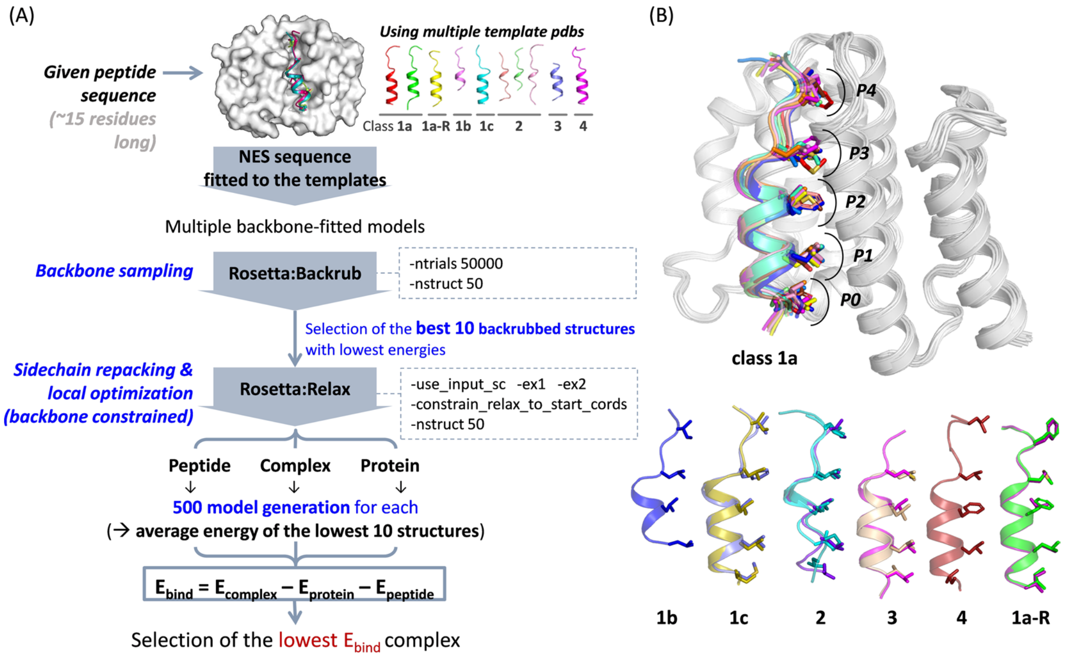

(A)
Given peptide sequence
(~15 residues long)
Using multiple template pdbs
Class 1a 1a-R 1b 1c 2 3 4
NES sequence fitted to the templates
Multiple backbone-fitted models
Backbone sampling
Rosetta:Backrub
-ntrials 50000
-nstruct 50
Selection of the best 10 backrubbed structures with lowest energies
Sidechain repacking & local optimization (backbone constrained)
Rosetta:Relax
-use_input_sc -ex1 -ex2
-constrain_relax_to_start_cords
-nstruct 50
Peptide Complex Protein
500 model generation for each
(→ average energy of the lowest 10 structures)
E_bind = E_complex – E_protein – E_peptide
Selection of the lowest E_bind complex
(B)
P4
P3
P2
P1
P0
class 1a
1b
1c
2
3
4
1a-R

Figure 5. Structure-based prediction of the stability of CRM1-NES peptide complex. (**A**) CRM1-NES peptide complex model generation and E_{bind} calculation procedure. (**B**) Generated models for the complex structures of CRM1-NES peptides with lowest E_{bind}. Class 1a peptides are displayed with CRM1 (in white) at the top. The hydrophobic (Φ) residues of these NES peptides (shown in the sticks) occupy the corresponding hydrophobic pockets (P0-P4) in CRM1. Peptides of other classes are shown at the bottom with the hydrophobic residues shown in the sticks.

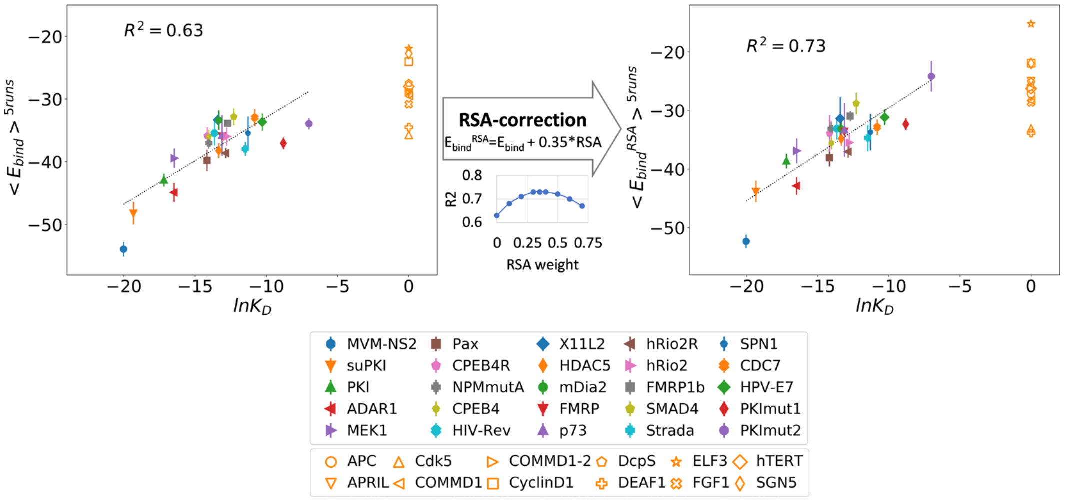


Figure 6. Correlation between the binding energies and the experimental K_D values. The binding scores are averaged in the five independent runs ($<E_{bind}>^{5runs}$; $<E_{bind}^{RSA}>^{5runs}$ for the RSA-corrected values) and compared to the logarithm of K_D values (lnK_D). The CRM1-binders with K_D values are shown in filled markers with error bars which are the standard deviation during the five runs. The false positives are shown in orange empty markers. In the middle, the correlation between R^2 and the weights for RSA during the E_{bind} correction is shown. The weight of 0.35 were applied for calculation of E_{bind}^{RSA}.

Possibility of non-binders to CRM1 among the NES-annotated regions. The databases like validNESs[15] and NESdb[16] provide valuable information on NES research, however, defining CRM1-dependent NES regions is still a difficult task. The expanding NES patterns result in many false positives. Also, the lack of information showing direct CRM1 binding to many annotated NES regions prevents development of accurate predictors using available data sets. Most published experimental studies were focused on showing that a protein is an export cargo, by deletion of the whole region encompassing a candidate NES or by mutation of all the suspected hydrophobic residue positions. These perturbations are drastic and may affect structural stability and result in defects of functions other than CRM1-binding and nuclear export. Therefore, one should interpret the

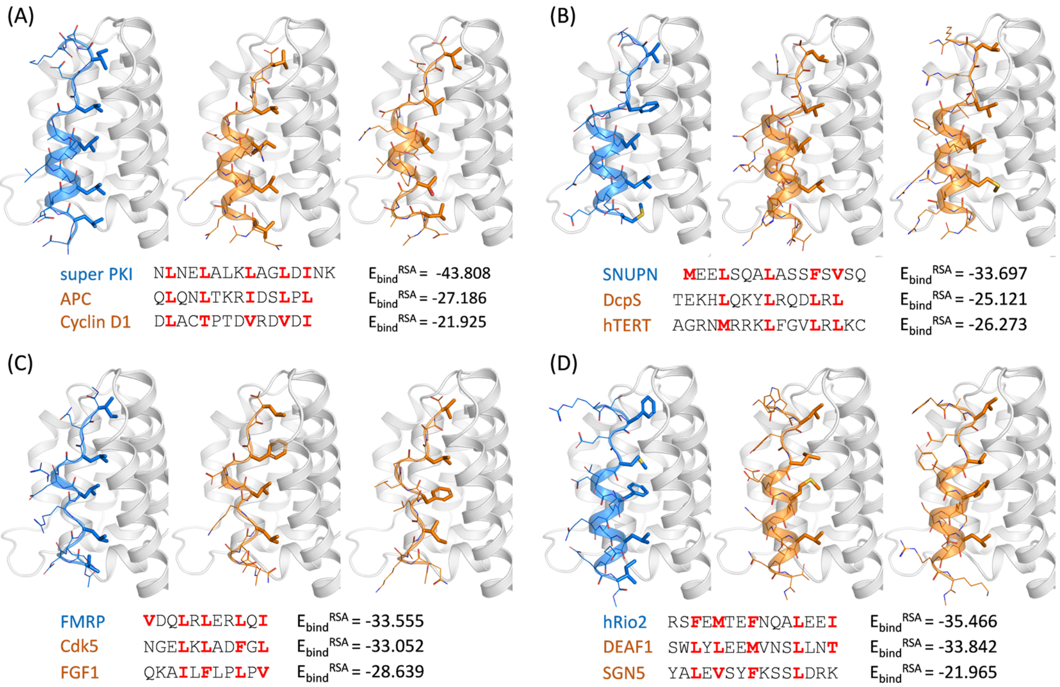

(A)
(B)
super PKI NLNELALKLAGLDINK E_{bind}^{RSA} = -43.808
APC QLQNLTKRIDSLPL E_{bind}^{RSA} = -27.186
Cyclin D1 DLACTPTDVRDVDI E_{bind}^{RSA} = -21.925
SNUPN MEELSQALASSFSVSQ E_{bind}^{RSA} = -33.697
DcpS TEKHLQKYLRQDLRL E_{bind}^{RSA} = -25.121
hTERT AGRNMRRKLFGVLRLKC E_{bind}^{RSA} = -26.273
(C)
(D)
FMRP VDQLRLERLQI E_{bind}^{RSA} = -33.555
Cdk5 NGELKLADFGL E_{bind}^{RSA} = -33.052
FGF1 QKAILFLPLPV E_{bind}^{RSA} = -28.639
hRio2 RSFEMTEFNQALEEI E_{bind}^{RSA} = -35.466
DEAF1 SWLYLEEMVNSLLNT E_{bind}^{RSA} = -33.842
SGN5 YALEVSYFKSSLDRK E_{bind}^{RSA} = -21.965

Figure 7. Comparison of the structural models and binding energies of the CRM1-binding NES motifs (blue) and false positive sequences (orange). CRM1 structure is colored in white. The hydrophobic residues are colored in red in the sequences and displayed as sticks in the structures. The spacer residues are represented as lines.

experimental data carefully to identify the CRM1-binding NES location, and it is always possible that regions which have been annotated as experimentally validated are not in fact functional NES motifs. Indeed, some of the annotated NES regions were found in the buried (highly ordered) protein domains (Fig. 4A,C). Some others can form β-strands in the middle of the segment (Fig. 4B,C) which would be rare in real NES sequences. Candidate segments that form β-strands and are located in the ordered region are observed in three cargoes including FAK ($_{91}$RSEE**V**HW**L**HVD**M**G**V**SS$_{106}$), MoKA ($_{190}$K**I**QT**L**H**L**VG**V**N**V**PE$_{203}$), and Sirt1 ($_{423}$DEVD**L**LIV**I**GSS**L**K**V**RP$_{239}$). We suggest that these segments have high possibility to be non-binders to CRM1 unless they unfold or transform their conformations upon specific conditions. Some cargo proteins might be exported following other events such as binding to an NES-containing adaptor protein.

Even if a segment fits the NES consensus and also satisfies the location criteria, these criteria are still not enough to locate the real NES segments in the whole protein sequence (see yellow highlighted segments in the online table). We tested the E_{bind} calculation to the all possible segments of the natural cargo proteins listed in Table 1. If a segment cannot form an energetically stable complex at the CRM1's NES binding groove, it is likely a non-binder to CRM1. As shown in Fig. 8, the NES candidates are likely to have the lower E_{bind} scores compared to other false positive segments. Among the seventeen cases, eleven cases have the NES candidate motifs with the lowest E_{bind}, and four cases have the NES regions with the second lowest E_{bind} but the difference between the lowest and second lowest is usually marginal (less than 2). Although the data set used in the structure-based modeling is quite small, the resulting binding energy values can discriminate between CRM1 binders and false positives. This structure-based prediction method can be utilized as one of the features to find real CRM1-dependent NES peptides in the pool of numerous false positive sequences.

Conclusion

In summary, we analyzed the structural prerequisites for CRM1-dependent NES motifs, i.e., accessibility (by locating disordered/ordered regions), adapting conformation (by predicting secondary structures), and the stability at the binding site (by applying structure-based modeling to calculate binding energies). The comprehensive table including all the possible consensus patterns with the disordered propensity plot, conserved domain information, and the predicted secondary structures provide valuable information for determining or correcting the most probable NES regions.

In light of the currently resolved crystal structures of CRM1-NES peptides with diverse classes, we modeled the CRM1-NES peptide complex structures and calculated the stability of the NES peptides at the CRM1 binding groove. The resulting binding energies correlate well to the experimental binding affinities, and we can distinguish the real NES motifs and false positives which both match NES consensus patterns. Also, we do not rely on the input sequence's pattern, rather use the energy function to select the most energetically favorable class template. Therefore, if the multiple patterns exist in one peptide segment, this energy calculation can be a tool to predict the peptide's conformation when it binds to CRM1. Although the method can still be improved, this study provides a starting point to predict NES motifs by combining sequence-based and structure-based approaches. Because our method is template-based modeling, it is difficult to adequately model NES motifs of classes other than those

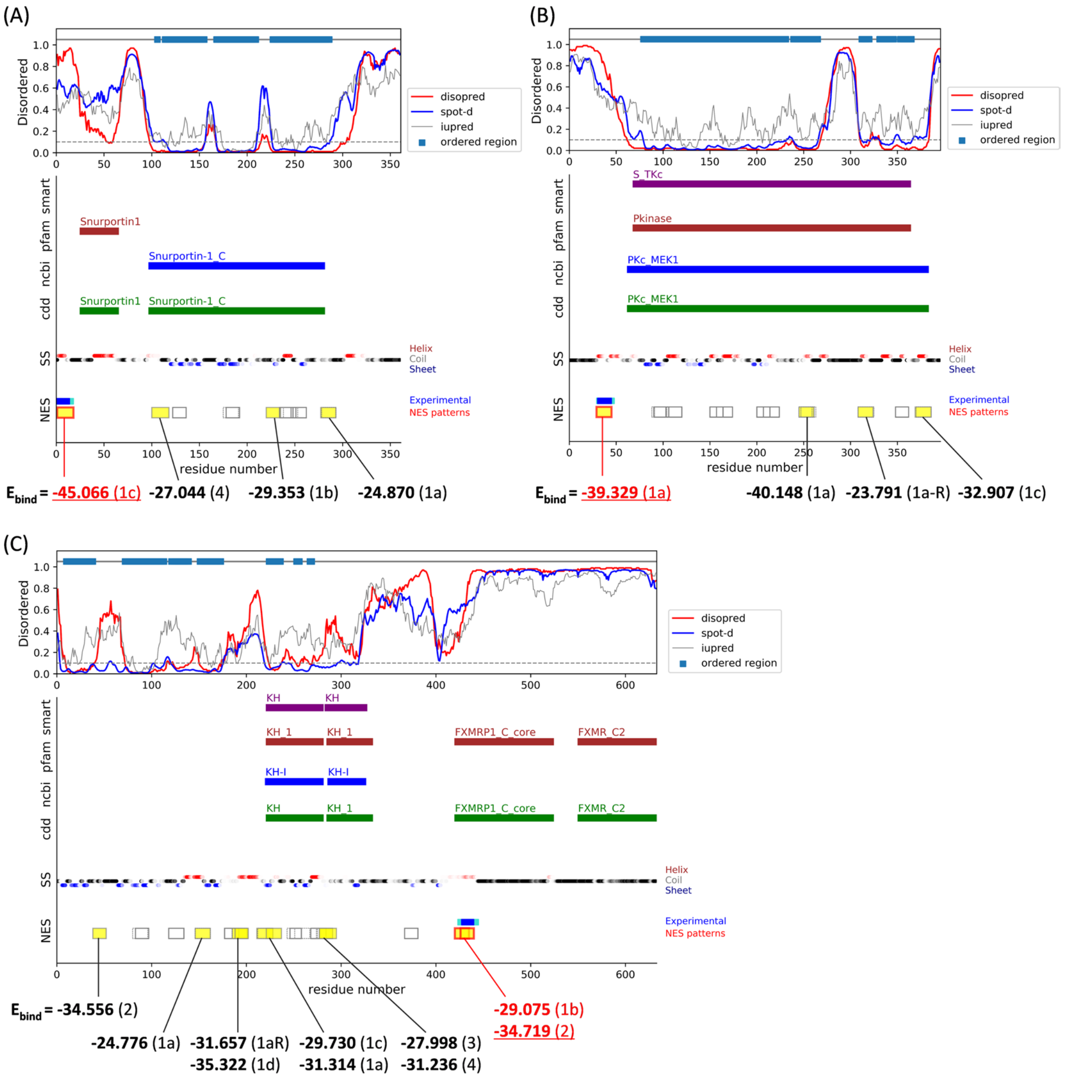


Figure 8. Distinguishing CRM1-binding NES motifs and false positives by E_{bind}. Location of NES consensus and their binding energies in (**A**) Snurportin-1 (O95149), (**B**) MEK1 (Q05116) and (**C**) FMRP (Q06787). The description for the plots is same as Fig. 2. The calculated E_{bind} scores for the important segments (pattern-matching segments which are not located in the highly ordered region and do not have β-strand conformation in the middle; yellow highlighted) were displayed together. The E_{bind} scores of the candidate NES motifs were underlined and marked in red. The classes of the consensus patterns are marked in parentheses.

of the templates. Since newly discovered NES motifs often deviate from the established consensus patterns, more structural information is definitely needed not only to understand new consensus patterns and NES-CRM1 binding mechanism but also to more accurately predict NES motifs.

Methods

Extraction of the NES consensus sequences. For the cargo proteins which have LMB sensitive data as CRM1-dependency annotated in NESdb[16] and validNES[15], the NES consensus-matching sequence segments were extracted by utilizing the modified version of the Kosugi consensus[16,20] (Fig. 2): $\Phi1\text{-}X_{1,2,3}\text{-}\Phi2\text{-}[\text{^}PW]_2\text{-}\Phi3\text{-}[\text{^}PW]\text{-}\Phi4$; $\Phi1\text{-}X_{2,3}\text{-}\Phi2\text{-}[\text{^}PW]_3\text{-}\Phi3\text{-}[\text{^}PW]\text{-}\Phi4$; or $\Phi1\text{-}X_2\text{-}\Phi2\text{-}X[\text{^}PW]_2\text{-}\Phi3\text{-}[\text{^}PW]_2\text{-}\Phi4$ ([^PW] is any of the 20 amino acids except Pro and Trp; Ala or Thr can be used only once at Φ1 or Φ2; X stands for any amino acid). If one segment or segments in the similar region (difference between the two segments' starting residue numbers <5) can be fitted to multiple patterns, all the possible patterns are recorded but prioritized based on the fact that: (i) the class 1a pattern is the most frequently observed class in the validated NES sets, suggesting that it interacts more preferentially with CRM1 than other classes[9,16,22]; (ii) in the current NES databases, class 3 sequences are as prevalent as NES motifs of classes 1c and 2[13]; (iii) the classes 1b and 1d can be found only in a few NES sequences, and the majority of the class 1d sequences can be overlapped to the class 1a pattern in the validated NES sets[9,13]; and (iv) reverse(−) of classes 3 and 4 appears to lack β-strands to hydrogen bond with the Lys residue and may not be ideal NES motifs[14]. This empirical class priority is defined as follows: (i) class 1a with five Φs (c1a-5) as priority 1; (ii) class 1 with four Φs (c1a-4), classes 1a-R, 2, 3, and 4 as priority 2; (iii) classes 1a/1c with Thr or Ala in one of their Φ1 or Φ2 positions as priority 3; (iv) classes 1b, 1d, 1c-reverse, and classes 2/3 with Thr or Ala in one of their Φ1 or Φ2 positions as priority 4, and (v) classes 1b/1d with Thr or Ala in one of their Φ1 or Φ2 positions as priority 5. The extracted regions are from the one residue before Φ0 to the two more residues after Φ4 (or shorter if located at the protein C- or N-termini). If the Φ2-Φ4 portion of the extracted region overlaps with experimental evidence (annotated as "mutations that affect nuclear export," "mutations that affect CRM1 binding," or "functional export signal" in NESdb, or annotated as "sites" in validNES), it is considered as a candidate NES. If not, it is deemed as a false positive.

Calculation of disorder propensity and definition of ordered regions. The disorder propensity of the cargo protein sequences is calculated using three different programs, DISOPRED3[25], SPOT-disorder[26], and IUPred2A[27]. For DISOPRED3 and SPOT-disorder calculation, which is based on multiple sequence alignment, uniref90_2015_01[34] database is used to find homologs during PSI-BLAST search[35]. In order to define ordered regions with high confidence, we applied strict cutoff values (~0.1) to decide the order/disorder border lines (note that the default values for disordered regions of these three programs here are ~0.5). If a residue's disorder propensities predicted by both DISOPRED and SPOT-disorder are below 0.1, the residue is defined as ordered ("O"). If not, the residue is recorded as potentially disordered ("D"). The predicted values by IUPred2A is also recorded for the reference. The sequence segment's location is determined by scanning the portion of "D" or "O" in the segment and flanking residues (20 residues at both sides) (Fig. S2A). 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. The segments in the boundary regions can be found at the end of the ordered regions, or they can locate in the ordered regions where some portions (>10%) have higher disorder propensity than the cutoff value.

Extraction of the conserved domain information of the cargo proteins. By using the Batch CD-search tool[36], the conserved domain information for the cargo protein sequences 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 expect value threshold of 0.01. The results were retrieved by the Concise mode.

Prediction of secondary structure. Secondary structures of the cargo protein sequences are predicted by PSIPRED Version 3.21[37]. During PSI-BLAST search[35] to find homologs, uniref90_2015_01[34] database is used. In the online table, the confidence level of the prediction is also colored by a gradient from dark (high confidence) to light (low confidence).

Relative binding energy (E_{bind}) prediction. Ten crystal structures of CRM1 bound to various NES peptides, including MVM-NS2 (PDB ID: 6CIT[31]), super PKI (unpublished data), FMRP-1b (5UWO[14]), SNUPN (3GB8[12]), FMRP (5UWJ[14]), SMAD4 (5UWU[14]), HIV-Rev (3NBZ[11]), X11L2 (5UWS[14]), and CPEB4 (5DIF[13]), were utilized as templates. For the CRM1 part, we extracted the residues from 479 to 655 (numbered in scCRM1) to reduce the computation time. For potential NES peptides, the positions from Φ0−1 to Φ4 + 2 positions were modeled (or a shorter segment in case a sequence used in the experimental K_D measure is shorter). A given peptide sequence is fitted to the backbone coordinates of every template structure. By using the Rosetta backrub module[38], the backbone conformations of the fitted NES peptide and the surrounding helices in CRM1 are sampled to generate 50 models (50,000 backrub Monte Carlo trials/steps were run for each model). Among them, five complex structures with the lowest energy are selected and then optimized by the Rosetta relax module[39,40], which searches the local conformational space around the starting structure. The relaxation was carried out 50 times for each model (i.e., the total number of models for a given peptide sequence is 10 × 50 = 500 models) with '-use_input_sc -ex1 -ex2' flag for more rigorous search. The backrub-modeled backbone conformation was constrained during the relaxation by applying '-constrain_relax_to_start_cords' flag. Structures of the CRM1 protein itself and the free peptide are also modeled separately with the same process. The all-atom energy function REF15 in Rosetta v.3.9 were utilized for all calculation.

The binding energy (E_{bind}) is calculated as $E_{complex} - E_{protein} - E_{peptide}$. The values for $E_{complex}$, $E_{protein}$, and $E_{peptide}$ are the average of the lowest 10 energy values among the 500 models. For $E_{peptide}$, we utilized the lowest $E_{peptide}$ among the all different backbone fitted models. Among the various template-fitted models, the one with the lowest E_{bind} score is selected. The E_{bind} scores were corrected with a solvent accessibility term calculated by the NACESS v.2.1.1 program[32], which calculates the atomic accessible surface defined by rolling a probe of given size around a vdw surface. To penalize the cavity at the interface of CRM1 and low-affinity binders (such as PKI double mutant), the RSA values for the hydrophobic residues at the interface (Fig. S3) were extracted and added to the E_{bind} scores with the optimized weight.

Data Availability

The datasets generated during and/or analyzed during the current study are included in this published article and available via: http://prodata.swmed.edu/nes_pattern_location/.

References

1. Fornerod, M., Ohno, M., Yoshida, M. & Mattaj, I. W. CRM1 is an export receptor for leucine-rich nuclear export signals. *Cell* **90**, 1051–1060, https://doi.org/10.1016/S0092-8674(00)80371-2 (1997).
2. Fukuda, M. *et al.* CRM1 is responsible for intracellular transport mediated by the nuclear export signal. *Nature* **390**, 308–311 (1997).
3. OssarehNazari, B., Bachelerie, F. & Dargemont, C. Evidence for a role of CRM1 in signal-mediated nuclear protein export. *Science* **278**, 141–144, https://doi.org/10.1126/science.278.5335.141 (1997).

4. Dickmanns, A., Monecke, T. & Ficner, R. Structural Basis of Targeting the Exportin CRM1 in Cancer. *Cells-Basel* **4**, 538–568, https://doi.org/10.3390/cells4030538 (2015).
5. Kau, T. R., Way, J. C. & Silver, P. A. Nuclear transport and cancer: From mechanism to intervention. *Nat Rev Cancer* **4**, 106–117, https://doi.org/10.1038/nrc1274 (2004).
6. Fischer, U., Huber, J., Boelens, W. C., Mattaj, I. W. & Luhrmann, R. The Hiv-1 Rev Activation Domain Is a Nuclear Export Signal That Accesses an Export Pathway Used by Specific Cellular Rnas. *Cell* **82**, 475–483, https://doi.org/10.1016/0092-8674(95)90436-0 (1995).
7. Wen, W., Meinkoth, J. L., Tsien, R. Y. & Taylor, S. S. Identification of a Signal for Rapid Export of Proteins from the Nucleus. *Cell* **82**, 463–473, https://doi.org/10.1016/0092-8674(95)90435-2 (1995).
8. la Cour, T. *et al*. Analysis and prediction of leucine-rich nuclear export signals. *Protein Eng Des Sel* **17**, 527–536, https://doi.org/10.1093/protein/gzh062 (2004).
9. Kosugi, S., Hasebe, M., Tomita, M. & Yanagawa, H. Nuclear Export Signal Consensus Sequences Defined Using a Localization-Based Yeast Selection System. *Traffic* **9**, 2053–2062, https://doi.org/10.1111/j.1600-0854.2008.00825.x (2008).
10. Monecke, T. *et al*. Crystal Structure of the Nuclear Export Receptor CRM1 in Complex with Snurportin1 and RanGTP. *Science* **324**, 1087–1091, https://doi.org/10.1126/science.1173388 (2009).
11. Guttler, T. *et al*. NES consensus redefined by structures of PKI-type and Rev-type nuclear export signals bound to CRM1. *Nat Struct Mol Biol* **17**, 1367–U1229, https://doi.org/10.1038/nsmb.1931 (2010).
12. Dong, X. H. *et al*. Structural basis for leucine-rich nuclear export signal recognition by CRM1. *Nature* **458**, 1136–U1171, https://doi.org/10.1038/nature07975 (2009).
13. Fung, H. Y. J., Fu, S. C., Brautigam, C. A. & Chook, Y. M. Structural determinants of nuclear export signal orientation in binding to exportin CRM1. *Elife* **4**, e10034, https://doi.org/10.7554/eLife.10034 (2015).
14. Fung, H. Y. J., Fu, S. C. & Chook, Y. M. Nuclear export receptor CRM1 recognizes diverse conformations in nuclear export signals. *Elife* **6**, e23961, https://doi.org/10.7554/eLife.23961 (2017).
15. Fu, S. C., Huang, H. C., Horton, P. & Juan, H. F. ValidNESs: a database of validated leucine-rich nuclear export signals. *Nucleic Acids Res* **41**, D338–D343, https://doi.org/10.1093/nar/gks936 (2013).
16. Xu, D. R., Grishin, N. V. & Chook, Y. M. NESdb: a database of NES-containing CRM1 cargoes. *Mol Biol Cell* **23**, 3673–3676, https://doi.org/10.1091/mbc.E12-01-0045 (2012).
17. Kirli, K. *et al*. A deep proteomics perspective on CRM1-mediated nuclear export and nucleocytoplasmic partitioning. *Elife* **4**, e11466, https://doi.org/10.7554/eLife.11466 (2015).
18. Fu, S. C., Imai, K. & Horton, P. Prediction of leucine-rich nuclear export signal containing proteins with NESsential. *Nucleic Acids Res* **39**, e111, https://doi.org/10.1093/nar/gkr493 (2011).
19. Kosugi, S., Yanagawa, H., Terauchi, R. & Tabata, S. NESmapper: Accurate Prediction of Leucine-Rich Nuclear Export Signals Using Activity-Based Profiles. *Plos Comput Biol* **10**, e1003841, https://doi.org/10.1371/journal.pcbi.1003841 (2014).
20. Xu, D. R. *et al*. LocNES: a computational tool for locating classical NESs in CRM1 cargo proteins. *Bioinformatics* **31**, 1357–1365, https://doi.org/10.1093/bioinformatics/btu826 (2015).
21. Prieto, G., Fullaondo, A. & Rodriguez, J. A. Prediction of nuclear export signals using weighted regular expressions (Wregex). *Bioinformatics* **30**, 1220–1227, https://doi.org/10.1093/bioinformatics/btu016 (2014).
22. Liku, M. E., Legere, E. A. & Moses, A. M. NoLogo: a new statistical model highlights the diversity and suggests new classes of Crm1-dependent nuclear export signals. *Bmc Bioinformatics* **19**, 65, https://doi.org/10.1186/s12859-018-2076-7 (2018).
23. Jehl, P., Manguy, J., Shields, D. C., Higgins, D. G. & Davey, N. E. ProViz-a web-based visualization tool to investigate the functional and evolutionary features of protein sequences. *Nucleic Acids Res* **44**, W11–W15, https://doi.org/10.1093/nar/gkw265 (2016).
24. Bienert, S. *et al*. The SWISS-MODEL Repository-new features and functionality. *Nucleic Acids Res* **45**, D313–D319, https://doi.org/10.1093/nar/gkw1132 (2017).
25. Jones, D. T. & Cozzetto, D. DISOPRED3: precise disordered region predictions with annotated protein-binding activity. *Bioinformatics* **31**, 857–863, https://doi.org/10.1093/bioinformatics/btu744 (2015).
26. Hanson, J., Yang, Y. D., Paliwal, K. & Zhou, Y. Q. Improving protein disorder prediction by deep bidirectional long short-term memory recurrent neural networks. *Bioinformatics* **33**, 685–692, https://doi.org/10.1093/bioinformatics/btw678 (2017).
27. Meszaros, B., Erdos, G. & Dosztanyi, Z. IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic Acids Res* **46**, W329–W337, https://doi.org/10.1093/nar/gky384 (2018).
28. Kim, J. Y. *et al*. HDAC1 nuclear export induced by pathological conditions is essential for the onset of axonal damage. *Nat Neurosci* **13**, 180–U163, https://doi.org/10.1038/nn.2471 (2010).
29. Bolli, N. *et al*. Born to be exported: COOH-terminal nuclear export signals of different strength ensure cytoplasmic accumulation of nucleophosmin leukemic mutants. *Cancer Res* **67**, 6230–6237, https://doi.org/10.1158/0008-5472.Can-07-0273 (2007).
30. Pinarbasi, E. S. *et al*. Active nuclear import and passive nuclear export are the primary determinants of TDP-43 localization. *Sci Rep-Uk* **8**, 7083, https://doi.org/10.1038/s41598-018-25008-4 (2018).
31. Fu, S. C., Fung, H. Y. J., Cagatay, T., Baumhardt, J. & Chook, Y. M. Correlation of CRM1-NES affinity with nuclear export activity. *Mol Biol Cell* **29**, 2037–2044, https://doi.org/10.1091/mbc.E18-02-0096 (2018).
32. 'NACCESS', computer program. (Department of Biochemistry and Molecular Biology, University College, London, 1993).
33. Xu, D. R., Farmer, A., Collett, G., Grishin, N. V. & Chook, Y. M. Sequence and structural analyses of nuclear export signals in the NESdb database. *Mol Biol Cell* **23**, 3677–3693, https://doi.org/10.1091/mbc.E12-01-0046 (2012).
34. Suzek, B. E. *et al*. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* **31**, 926–932, https://doi.org/10.1093/bioinformatics/btu739 (2015).
35. Altschul, S. F. *et al*. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389–3402, https://doi.org/10.1093/nar/25.17.3389 (1997).
36. Marchler-Bauer, A. & Bryant, S. H. CD-Search: protein domain annotations on the fly. *Nucleic Acids Res* **32**, W327–W331, https://doi.org/10.1093/nar/gkh454 (2004).
37. Jones, D. T. Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol* **292**, 195–202, https://doi.org/10.1006/jmbi.1999.3091 (1999).
38. Smith, C. A. & Kortemme, T. Backrub-like backbone simulation recapitulates natural protein conformational variability and improves mutant side-chain prediction. *J Mol Biol* **380**, 742–756, https://doi.org/10.1016/j.jmb.2008.05.023 (2008).
39. Nivon, L. G., Moretti, R. & Baker, D. A Pareto-Optimal Refinement Method for Protein Design Scaffolds. *Plos One* **8**, e59004, https://doi.org/10.1371/journal.pone.0059004 (2013).
40. Conway, P., Tyka, M. D., DiMaio, F., Konerding, D. E. & Baker, D. Relaxation of backbone bond geometry improves protein energy landscape modeling. *Protein Sci* **23**, 47–55, https://doi.org/10.1002/pro.2389 (2014).

Acknowledgements

This work is 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). 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.

Author Contributions

N.V.G. conceived of the presented idea and designed the research. Y.L. and J.P. developed the theory, performed the simulation, and analyzed the data. J.M.B. and Y.M.C. performed the experimental validation of the binding affinities and provided the structural data. Y.L. wrote the manuscript. Y.L., J.P., J.M.B., Y.M.C. and N.V.G. contributed to the interpretation of the results and revised the manuscript. N.V.G. supervised all the study.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-019-43004-0.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.