**Protein Engineering, Design & Selection** vol. 17 no. 6 pp. 527–536, 2004
Published online August 16, 2004 doi:10.1093/protein/gzh062

# Analysis and prediction of leucine-rich nuclear export signals

**Tanja la Cour[1,2], Lars Kiemer[1,2], Anne Mølgaard[1], Ramneek Gupta[1], Karen Skriver[3] and Søren Brunak[1,4]**

[1]Center for Biological Sequence Analysis, Biocentrum-DTU, Technical University of Denmark, Building 208, DK-2800 Lyngby and [3]Department of Protein Chemistry, Institute of Molecular Biology, University of Copenhagen, Øster Farimagsgade 2A, DK-1353 Copenhagen K, Denmark

[2]These authors contributed equally to this work.

[4]To whom correspondence should be addressed.
E-mail: brunak@cbs.dtu.dk

**We present a thorough analysis of nuclear export signals and a prediction server, which we have made publicly available. The machine learning prediction method is a significant improvement over the generally used consensus patterns. Nuclear export signals (NESs) are extremely important regulators of the subcellular location of proteins. This regulation has an impact on transcription and other nuclear processes, which are fundamental to the viability of the cell. NESs are studied in relation to cancer, the cell cycle, cell differentiation and other important aspects of molecular biology. Our conclusion from this analysis is that the most important properties of NESs are accessibility and flexibility allowing relevant proteins to interact with the signal. Furthermore, we show that not only the known hydrophobic residues are important in defining a nuclear export signals. We employ both neural networks and hidden Markov models in the prediction algorithm and verify the method on the most recently discovered NESs. The NES predictor (NetNES) is made available for general use at http://www.cbs.dtu.dk/.**
*Keywords*: analysis/NES/nuclear export signals/prediction/structure

## Introduction

Eukaryotic cells are characterized by having their genetic material enclosed in a special compartment—the nucleus. Thereby, transcriptional and translational events are physically separated in eukaryotic cells. Compartmentalization protects the genetic material and offers additional levels of transcriptional and translational regulation at the price of energy-consuming regulated transport of macromolecules.

Transport across the nuclear envelope occurs through the structurally conserved nuclear pore complex (NPC), a huge proteinaceous structure forming an aqueous channel (Stoffler *et al*., 1999; Ryan and Wente, 2000). Even though the internal diameter of the NPC allows passive diffusion of small proteins (<60 kDa), most proteins with nuclear functions appear to be actively transported in and out of the nucleus. Active nucleocytoplasmic transport is a signal-dependent process mostly mediated by a family of homologous transport receptors belonging to the importin family, also called karyopherin receptors or importins/exportins (Strom and Weis, 2001).

Several pathways of nuclear export have been identified (Ossareh-Nazari *et al*., 2001), but the export pathway, for which most knowledge has accumulated so far, requires a leucine-rich nuclear export signal (NES). This signal was first discovered in the human immunodeficiency virus type 1 (HIV-1) Rev protein (Fischer *et al*., 1995) and cAMP-dependent protein kinase inhibitor (PKI) (Wen *et al*., 1995). The karyopherin receptor CRM1/exportin 1/XPO1, hereafter called CRM1, has been identified as the export receptor for leucine-rich NESs in several organisms (Fornerod *et al*., 1997; Fukuda *et al*., 1997; Neville *et al*., 1997; Ossareh-Nazari *et al*., 1997; Stade *et al*., 1997; Haasen *et al*., 1999) and is an evolutionarily conserved protein typically having 42–50% sequence identity between orthologs from *Homo sapiens, Saccharomyces cerevisiae* and *Arabidopsis thaliana* (Haasen *et al*., 1999). CRM1-mediated export is effectively inhibited by the fungicide leptomycin B (LMB) (Fornerod *et al*., 1997), providing excellent experimental verification of this pathway. LMB has been shown to bind covalently to a cysteine residue in CRM1 (Kudo *et al*., 1999a) except in the yeast CRM1 ortholog (Crm1p), which as an exception to the rule is LMB-insensitive in its wild-type form, but becomes LMB-sensitive upon introduction of a T539C point mutation (Neville and Rosbash, 1999).

There is some controversy as to whether CRM1 interacts directly with leucine-rich NESs in the absence of RanGTP (Fukuda *et al*., 1997; Ossareh-Nazari *et al*., 1997; Ossareh-Nazari and Dargemont, 1999) or only in the presence of RanGTP (Askjaer *et al*., 1998). Nevertheless, there is general agreement about RanGTP and leucine-rich NESs binding cooperatively to CRM1 upon formation of a ternary CRM1–RanGTP–NES complex (Fornerod *et al*., 1997; Askjaer *et al*., 1998; Ossareh-Nazari and Dargemont, 1999). Furthermore, RanBP3/Yrb2p has been shown to function as an essential export cofactor (Taura *et al*., 1998), that enhances the affinity of CRM1 for RanGTP and thereby also for the NES substrate (Englmeier *et al*., 2001; Lindsay *et al*., 2001). Following export, several proteins (e.g. RanBP1, RanGAP and Nup358/RanBP2) are involved in the disassembly of the export complex and recycling of CRM1 (Bischoff *et al*., 1995; Bischoff and Gorlich, 1997; Bernad *et al*., 2004).

Not all NES substrates are constitutively exported from the nucleus, meaning that CRM1-mediated export is a regulated event. Several ways of regulating NES-dependent export have been reported. These include masking/unmasking of NESs (Li *et al*., 1998; Stommel *et al*., 1999; Seimiya *et al*., 2000; Heerklotz *et al*., 2001; Kobayashi *et al*., 2001; Craig *et al*., 2002), phosphorylation (Engel *et al*., 1998; Ohno *et al*., 2000; McKinsey *et al*., 2001; Zhang and Xiong, 2001; Brunet *et al*., 2002) and even disulfide bond formation as a result of oxidation (Yan *et al*., 1998; Kudo *et al*., 1999b; Kuge *et al*., 2001). Although so far speculative, the availability of specific export cofactors might also participate in the export regulation of specific NES substrates.

Besides being important for a better understanding of eukaryotic gene regulation, insight into the mechanism and regulation of nuclear export might also be relevant from a therapeutic point of view: Many of the reported nucleocytoplasmic shuttle proteins are involved in signal transduction events and cell cycle regulation (Gama-Carvalho and Carmo-Fonseca, 2001; Kau and Silver, 2003). In fact, the publicly available database of NES proteins, NESbase 1.0 (la Cour *et al*., 2003), contains 17.3% tumor suppressors and oncoproteins and the misregulation of their subcellular localization has been shown to be involved in the development of different cancers (Fabbro and Henderson, 2003; Kau *et al*., 2004). In addition, export of unspliced and partially spliced HIV-1 mRNA depends on the leucine-rich NES of the HIV-1 Rev protein (Hope, 1999).

To date, leucine-rich NESs have been reported on a case-by-case basis and have largely been identified using consensus patterns followed by experimental validation. We have compiled experimentally verified leucine-rich NESs in the database NESbase 1.0 (la Cour *et al*., 2003), in order to investigate the sequence and structural requirements of leucine-rich NESs responsible for their specific interaction with CRM1 and to develop a good NES prediction method as no such method is currently available. Prediction is a valuable tool for the experimentalist as a reliable prediction can save many hours of laboratory effort and thereby increase the speed of new discoveries and expand the current knowledge on CRM1-mediated nuclear export.

A functionally related motif on which considerable work has been done in developing computational methods is nuclear localization signals (NLSs). In contrast to the nuclear export signal, this motif is responsible for nuclear import. It has been shown that a collection of patterns consisting of known NLS motifs and cleverly *in silico*-mutated motifs can be used to identify most known signals (Cokol *et al*., 2000; Nair *et al*., 2003). However, as our work shows, identifying NESs is significantly more complicated and consensus patterns unfortunately do not suffice.

Neural networks and hidden Markov models have been successfully employed in similar biological problems numerous times (Durbin *et al*., 1998; Baldi and Brunak, 2001). Our goal with this work was to collect available information about NESs and to use this information together with bioinformatic analyses in order to make a thorough examination and characterization of the signal and a prediction method useful for the community. Besides being a help to individual researchers working with NESs, reliable predictions of NESs would be of great advantage in the immense task of proteome annotation that lies ahead in the wake of the genome sequencing projects.

## Materials and methods

### *Data set*

NESbase 1.0 (la Cour *et al*., 2003) contains 75 entries with information on CRM1 dependency and whether or not each NES has been shown to be sufficient and/or necessary for export. The data set used in this study comprises 64 proteins containing 67 high-confidence nuclear export signals.

### *Sequence logos*

The height of the amino acid one-letter abbreviations reflects the Shannon information content (Shannon, 1948) in units of bits at that specific position in the multiple sequence alignment (Schneider and Stephens, 1990).

### *Structure retrieval and visualization*

PDB was searched for structural hits to all 67 NESs included in this study using the as yet unpublished, in-house tool GETSTRUCT program (A.Mølgaard). If more than one structural hit was obtained, sequence identity and structure resolution were considered when selecting the structure to use.

Structures were visualized using the programs MOLSCRIPT (Kraulis, 1991) and RASTER3D (Merrit and Bacon, 1997). The coloring by temperature factor was based on the mean and standard deviation of temperature factors in the individual structures, such that temperature factors lower than the mean minus standard deviation are colored blue and temperature factors higher than mean plus standard deviation are colored red. In that manner, temperature factors within the standard deviation are purple.

### *Cross-validation and redundancy reduction*

Based on NESbase (la Cour *et al*., 2003) a redundancy-reduced, homology-partitioned four-fold cross-validation data set was constructed (Hobohm *et al*., 1992). For negative examples the remaining residues of the NES-containing sequences were used, except in cases where the NES of the sequence in question had not been shown to be crucial for nuclear export (i.e. indicating that the sequence could contain another NES). Generally, only NES sequences proved to be functional upon introduction in other proteins were used. Each of the four test sets thus contained 16 or 17 positive examples and of the order of 6000–8000 negative examples. These data sets were used for training of both the hidden Markov model and the neural network and all testing results reported represent combined values of the four test sets, which are run individually with four separate neural networks or hidden Markov models to avoid testing on sequences included in training sets.

### *Construction of the hidden Markov model*

The hidden Markov model was trained and decoded with an unpublished, in-house tool, ANHMM v1.211 (A.Krogh). The model consisted of four states representing hydrophobic positions with flexible states in between, making it capable of adapting to a variable number of residues between the hydrophobic positions. The remainder of the sequence was modeled in a single state. The model was trained on labeled data in which the stretch of amino acids encompassing the hydrophobic residues was labeled 'NES' and the remainder of each sequences labeled 'not NES'. Outputs from the program are posterior probabilities of the labels used ('NES', 'not NES') and thus fall in the range [0;1].

### *Training the neural networks*

The artificial neural networks used in this work were of the standard feed-forward type. Sparse encoding was used for translating the amino acids to data input for the networks as described previously (Blom *et al*., 1996; Nielsen *et al*., 1997). The network obtained optimal performance using a three-layer network with two hidden neurons and a window size of 15 amino acids (corresponding to a true prediction on the window encompassing positions P-14 through P0 in Figure 1). Network

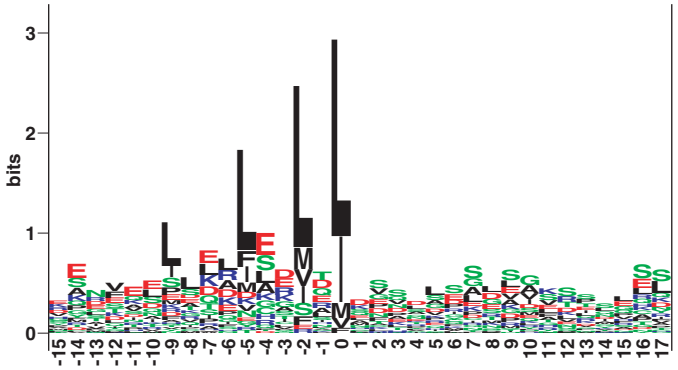


**Fig. 1.** Sequence logo of 67 experimentally verified NESs aligned by the hydrophobic anchor position at position P0.

performance was measured in terms of Matthews correlation coefficients (Matthews, 1975).

### *Integrating the models and predicting a signal*

When a sequence is submitted to the predictor, every residue is assigned a hidden Markov model score and a neural network score, both of which are an average of the scores from the four models or networks created during cross-validation. The post-processing algorithm then assigns a score to every sequence position (P), which is composed of an average of the neural network scores of the preceding four residues, the neural network score at P and the hidden Markov model score at position P. Both the neural network score and the hidden Markov model score are in the range 0–1.

To take advantage of the differences in the two models, neural network scores >0.5 are multiplied by 1.1 and scores <0.4 have 0.4 subtracted. Likewise, HMM scores <0.0001 are assigned −0.1 because the HMM almost never reports the value 0 in regions containing an NES. Finally, if the combined score is negative it is reported as zero, giving a theoretical NetNES scoring range of 0–2.1. This post-processing algorithm is designed to benefit from the fact that the HMM usually assigns very low values only to areas that are not NES regions and the neural network usually assigns high scores to at least some of the hydrophobic positions in the motif—if not all. The algorithm thus integrates the observed superior specificity of the hidden Markov model with the observed superior sensitivity of the neural network.

The web server output displays scores for the hidden Markov models, the neural networks and the post-processing score in order to present the user with all available information.

### *Performance evaluation and ROC curve*

To take into account the fact that an NES is not a precisely defined motif in terms of position like a proteinase cleavage site, we assigned a false positive only to above cut-off scores more than 15 amino acids from a true site. Furthermore, if several above cut-off scores were found within a 15 amino acid window, they were counted as only one positive.

The ROC (sensitivity/specificity) curve is based on a total number of negatives (277) which is much lower that the actual value (31 563). This was necessary for two reasons: (1) 96% of all sequence positions in the test sets are assigned 0 owing to the post-processing algorithm, (2) we evaluate a false positive in a 15 amino acid window context (see above). The total number of negatives was set to 277, corresponding to the number of positives (both true and false) obtained with a cut-off of 0.001. It must be noted, however, that this underestimates the performance of the model.

## Results

### *NES sequence logo*

To visualize sequence conservation of nuclear export signals, a sequence logo (Schneider and Stephens, 1990) was generated (Figure 1) from a multiple alignment of 67 high-confidence NESs. The NESs were aligned by the hydrophobic anchor position at position P0. This anchor position was assigned as the last hydrophobic residue in each NES signal, as there is some experimental indication of NES activity being more susceptible to mutation of the hydrophobic residues in the C-terminal end of the signal than in the N-terminal end (Wen *et al*., 1995; Kudo *et al*., 1999b). Therefore, the conservation of a large hydrophobic residue at position P0 in the logo has been incorporated automatically and does not provide additional evidence of this hydrophobic residue being the most conserved of the hydrophobic residues comprising the signal.

The most conserved pattern revealed by the sequence logo is the LxxLxL motif comprised of the last three hydrophobic residues in the signal, whereas the position of the first hydrophobic residue is less conserved. Also evident is the prevalence of glutamate, aspartate and serine: Any of these amino acids ranges in the top at nearly every position not occupied by hydrophobic residues in the NES region (Figure 1). This indicates a preference for these amino acids in the region, although a pattern is not evident.

### *Promiscuity and spacing of hydrophobic residues*

Almost two-thirds (63%) of the 67 NESs used in this study deviate from the generally accepted consensus L–x(2,3)–[LIVFM]–x(2,3)–L–x–[LI] (Bogerd *et al*., 1996), comprised of four conserved large hydrophobic residues—primarily leucines—with a variable spacing in-between them. However, although we identify only 25 NESs with the consensus pattern, we retrieve 23 other, random occurrences (false positives) in the 64 sequences containing the 67 signals. Present in this data set is one common deviation from the generally accepted consensus, namely the higher degree of promiscuity with respect to the specific hydrophobic residues for the last two conserved hydrophobic residues in the signal: 88% of the signals in the data set have [LIVFM]–x–[LIVFM] in the end of the signal, whereas only 58% have L–x–[LI]. Promiscuity of the first hydrophobic residue is also revealed since 72% of the NESs is represented by [LIVFM]–x(2,3)–[LIVFM]–x(2,3)–[LIVFM]–x–[LIVFM], whereas only 52% are detected if requiring a leucine at the first position. Unfortunately, as sensitivity rises so does the number of false positives: with the more slack pattern above identifying 72% of the NESs we also find 252 false positives in the data set. In general, the more specific the pattern, the fewer false NES signals would be expected, but a more specific pattern fits fewer NESs.

Another variable feature of nuclear export signals is the spacing between the hydrophobic amino acid residues. The sequence logo shows the most conserved pattern to be LxxLxL, where L can be either L, I, V, F or M. In fact, 75% of the NESs in the data set fit an LxxLxL pattern, whereas only 11% fit an LxxxLxL pattern while not fitting LxxLxL. Still, 15% of the

NESs do not conform to either of the patterns (LxxLxL or LxxxLxL).

These results clearly indicate that the signal is complex and involves several sequence positions with different amino acid distributions. It also demonstrates the limitations of patterns, especially in dealing with complex, promiscuous signals.

### *NES regions are rich in glutamate, aspartate and serine*

Analysis of the amino acid composition of different segments of the NES region and comparison with the overall amino acid composition of the proteins in the data set revealed statistically significant over-representation of glutamate, aspartate, serine and leucine. In Figure 2, the frequency of a given amino acid at a given position is compared with the baseline level of that particular amino acid in the data set. Each of these four amino acids is over-represented in discrete segments of the NES region. The individual amino acid over-representations are statistically significant, as verified by a hypergeometric test (Table I).

### *The majority of NES regions are negatively charged*

The over-representation of acidic residues in NES regions implies that NESs could have an overall negative charge. In Figure 3, the distribution of isoelectric points (p*I*) for two different sequence segments covering the NES signal and for full-length NES protein are compared, revealing a trend of NES regions having lower p*I* than the full-length proteins. The NES region P-15 to P10 covers the segment of sequence with the highest over-representation of glutamate and aspartate, whereas NES region P-15 to P17 also includes the segment of serine over-representation. As expected, more nuclear export signals therefore have low p*I* in NES region P-15 to P10 than P-15 to P17. In fact, in NES region P-15 to P10 more than half of the NESs have a p*I* between 4 and 5 and >80% have a p*I* <6.27. Curiously, 17% the NESs have p*I* >8. In conclusion, most NESs are charged and of these most are negatively charged.

It is well known that charged regions are over-represented on the surface of proteins where they interact with the aqueous environment. Therefore, the prevalence of charged residues in the signal could serve to keep the NES region surface exposed.

### *Location of nuclear export signals in the tertiary structure of proteins*

What does the CRM1/exportin receptor recognize when interacting with NES regions? The Protein Data Bank (PDB) was searched for structurally determined NES proteins or close homologs using the as yet unpublished, in-house tool GETSTRUCT program (A.Mølgaard), a BLAST-based alignment

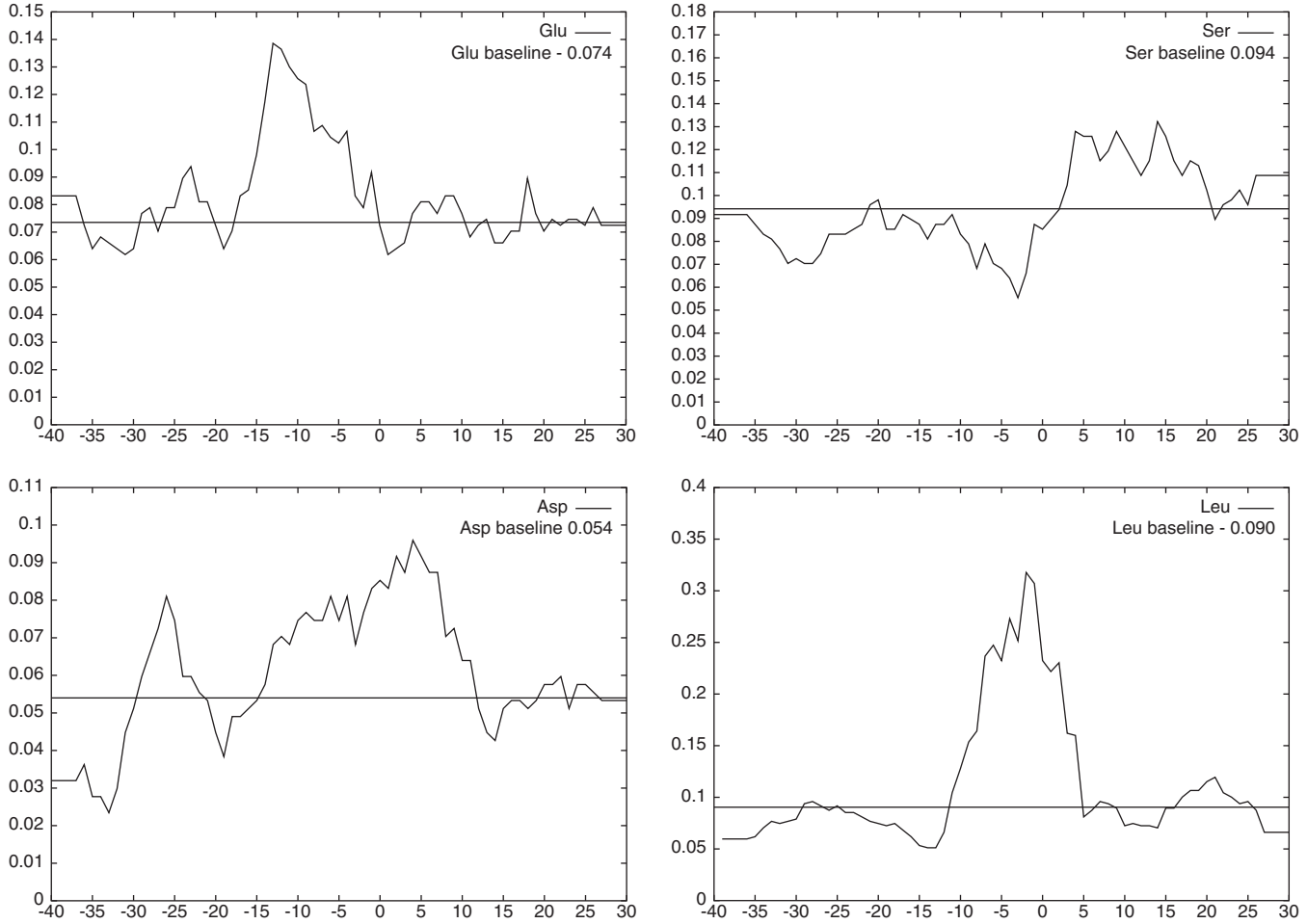


**Fig. 2.** Frequencies of the over-represented amino acids versus position. Position P0 corresponds to the hydrophobic anchor position. Also shown is the average frequency in the complete sequences. The curves are smoothed by plotting the average frequency over a window of seven residues (three to the right and three to the left of the position in question).

**Table I.** Counts and significance scores for over-represented amino acids in their respective segments

| Amino acid | Segment | Occurrences of amino acid within segment | | Sequences containing amino acid within segment | | Occurrences per sequence |
|---|---|---|---|---|---|---|
| | | Occurrences | Significance | Occurrences | Significance | |
| Glutamate | P-15–P-10 | 61 | 9.02 | 43 | 5.92 | 0.924 ± 0.882 |
| Aspartate | P-10–P10 | 116 | 5.26 | 56 | 1.57 | 1.731 ± 1.420 |
| Serine | P4–P17 | 116 | 3.59 | 54 | 1.72 | 1.841 ± 1.472 |
| Leucine | P-10–P0 | 211 | Inf | 67 | Inf | 3.149 ± 1.048 |

Statistical analysis of amino acid over-representation in the data set was performed and significance was estimated by a hypergeometric test. The over-representation is statistically significant when the significance score exceeds 1. Inf denotes a very high significance going towards infinity.

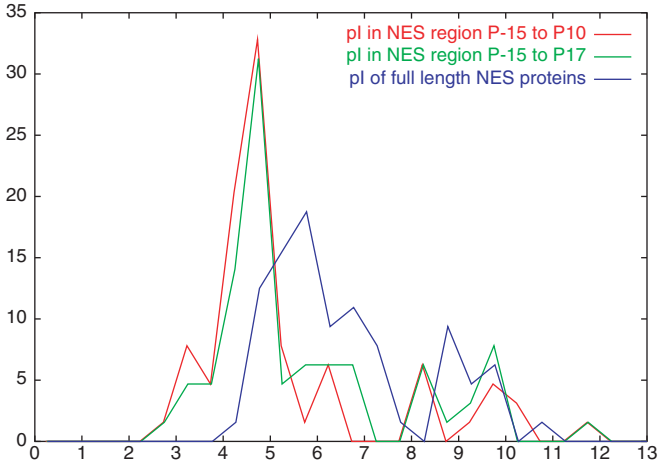


**Fig. 3.** The isoelectric point distribution of NESs. The number of sequences in a 0.5 p*I* range was plotted against the p*I*. Using the ProtParam tool at Expasy (Appel *et al*., 1994), p*I* was calculated for both full-length NES proteins and NES regions P-15–P10 and P-15–P17. NES region P-15–P10 includes only the region where either glutamate or aspartate is significantly over-represented. NES region P-15–P17 also includes the region of serine over-representation.

tool for finding protein structures in PDB. Eight proteins, six X-ray structures and two NMR structures were found, of which four were close homologs to the query proteins comprising nearly identical NESs. The sequences of the remaining four hits were identical with those of the query proteins (Table II).

Several of the NES regions are α-helical and packed up against another α-helix either in α-helical bundles or globular α-helical domains (Figure 4, available as Supplementary material at *PEDS* Online). The most C-terminal leucine in a NES tends to be exposed or lid covered and thereby available for interaction. Each atom in an X-ray structure is assigned a *B*-factor (the temperature factor), which is a measure of the dynamic and static disorder of any particular atom. The degree of exposure and flexibility of NES regions for the six X-ray structures is also visualized in Figure 4 in the Supplementary material. Most of the signals are either in very flexible regions or close to very flexible regions.

### *Experimentally determined secondary structure of nuclear export signals*

One of the eight structures of NES proteins obtained, actin from *S.cerevisiae* (1YAG), is reported to contain two NESs (Wada *et al*., 1998), which is why we have structural information for nine NESs altogether. Six of these nine NESs are α-helical and an additional two have a single turn of α-helix N-terminally (Figure 5, available as Supplementary material at *PEDS* Online). This is in agreement with the sequence logo (Figure 1), which with its pattern of hydrophobic residues and charged or polar residues suggests an amphipathic α-helical secondary structure. The only NES to contain no α-helical structure is the first of the two nuclear export signals found in yeast actin and this NES is furthermore one of the two signals containing α-strand. Despite these exceptions, there is a clear tendency for NESs to be α-helical in the N-terminal part of the signal based on the structural data available and, furthermore, for all of the hydrophobic residues to be protruding from one side of the helix. Prediction of protein secondary structure using PSIPRED (Jones, 1999) on the available data in NESbase predicts an α-helix in the N-terminal part of the signal (P-11 to P-4) in 70% of the NESs, coil in 23% and β-strand in only 6% (for comparison, the full-length sequences were predicted to contain 31% helix, 55% coil and 14% β-strand). This result indicates a strong preference for α-helix and bias against β-strand and coil. Also apparent from the structural data, the NES is located in highly flexible regions of the proteins, exceptions being the structures 1KWP and 1KHU (Figure 6, available as Supplementary material at *PEDS* Online).

### *Structural properties of flanking regions*

On inspection of the structure of the nuclear export signals, we noted that not only does the structure of the NES itself seem similar among the structures available, but also the structure of the flanking regions. The NES is usually located in α-helix, but close to the transition between two structural elements, which in most cases both are α-helices (Figure 7, available as Supplementary material at *PEDS* Online). The hydrophobic residues of the helix all protrude from the same side of the helix in which the signal is located. Again, as expected from an amphipatic α-helix, the polar residues present in the signal or close to the signal orient opposite to the hydrophobic residues. Six of the nine structures have one or more glutamate residues in the α-helix in which the NES hydrophobic residues are located (Figure 7, available as Supplementary material at *PEDS* Online).

### *Superposition of α-helical nuclear export signals*

The structural alignment of the six α-helical NESs confirms the notion of having three hydrophobic residues ‘stacking’ on one side of an α-helix (Figure 8). Curiously, these structures are best aligned in the N-terminal part of the signal including the first three conserved hydrophobic residues, but bend off differently in the

**Table II.** Available protein structures containing nuclear export signals

| PDB | Swiss-Prot | Identity | Res | *B*-factor | Reference |
|---|---|---|---|---|---|
| 1BF5 A | P42224 | 543/575 (94%) | 2.9 | 39.28 ± 22.03 | Chen *et al.* (1998) |
| 1DJX B | P10688 | 561/599 (93%) | 2.3 | 36.58 ± 18.47 | Essen *et al.* (1997) |
| 1KHU A | Q15797 | 198/198 (100%) | 2.5 | 44.65 ± 14.10 | Qin *et al.* (2001) |
| 1KWP A | P49138 | 312/340 (91%) | 2.8 | 43.01 ± 9.02 | Meng *et al.* (2002) |
| 1M5I A | P25054 | 105/110 (95%) | 2.0 | 38.46 ± 12.01 | Tickenbrock *et al.* (2002) |
| 1YAG A | P02579 | 372/372 (100%) | 1.9 | 21.12 ± 14.45 | Vorobiev *et al.* (2003) |
| 1JM7 A | P38398 | 103/103 (100%) | NMR | N/A | Brzovic *et al.* (2001) |
| 1OLG A | P04637 | 42/42 (100%) | NMR | N/A | Clore *et al.* (1994) |

Searching PDB yielded eight structures, which displayed high sequence identity with the Swiss-Prot query proteins (Identity). Some of the structures have missing atoms in the vicinity of the NES region, but if so, this is noted in the relevant figures. Average *B*-factors are calculated from C-α *B*-factors of all protein chains in the structure. Swiss-Prot refers to Swiss-Prot accession number of the query sequence; *Res* is the resolution in Å.

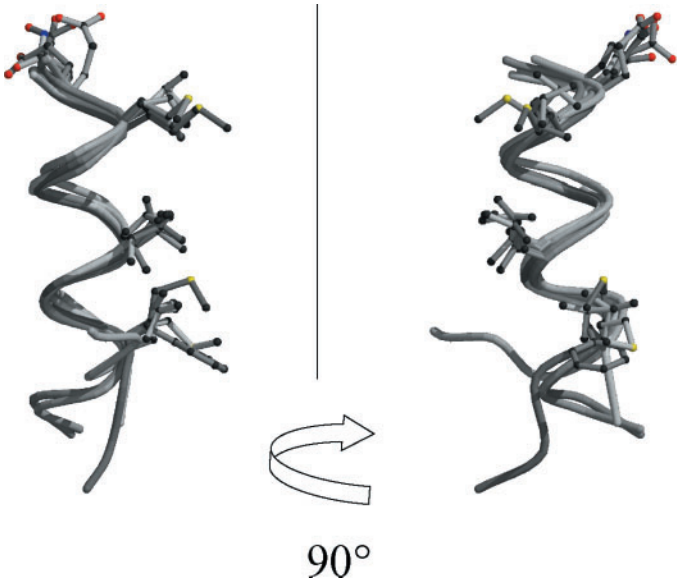


**Fig. 8.** Structural alignment of α-helical nuclear export signals. Structural alignment was performed using the program O (Jones *et al.*, 1991). The structure 1KWP was used as template and 1JM7, 1OLG, 1DJX, 1M5I and 1BF5 were aligned fitting the C-α coordinates of the hydrophobic residues. 1BF5, which has a divergent spacing of hydrophobic residues and is furthermore not visible in the entire region, is only aligned on the last two hydrophobic positions.

C-terminal end containing the last hydrophobic residue. The five structures that are aligned on three hydrophobic residues in Figure 8 share a common spacing of hydrophobic residues and four of these have a glutamate residue preceding the first hydrophobic residue of the signal, while the fifth has an asparagine residue at that position. This is not representative of the data set, however, since only 36 NESs (54%) have LxxxLxxL spacing of hydrophobic residues and of those only eight have a glutamate preceding the first hydrophobic residue.

To summarize the structural analyses of the available structures from PDB, we conclude that nuclear export signals reside in surface-exposed, highly flexible regions of the protein and usually form α-helical secondary structure.

### *Prediction of leucine-rich NESs by artificial neural networks and hidden Markov models*

Our initial approach towards NES recognition was neural networks trained with different window sizes. Networks with a layer of hidden units (i.e. non-linear networks) performed better than linear networks containing no hidden units (data not shown). This indicates a presence of non-linear correlations between amino acids in the NES regions, meaning that they do not contribute independently to NES activity. Neural networks encompassing from six up to 16 amino acids ending on the last hydrophobic position (P0) yielded predictive performance with Matthews correlation coefficients (Matthews, 1975) up to 0.35, which is somewhat low for a useful method. Networks with smaller window sizes centered on the individual, strongly conserved hydrophobic residues performed equally bad or worse. Combinations of a large window size network attempting to predict the entire NES region and the smaller networks centered on the hydrophobic residues raised the correlation coefficient to ~0.4.

We then tried to construct a hidden Markov model hoping that this approach would do better than the networks. However, again using 4-fold cross-validation a Matthews correlation coefficient of only ~0.4 was obtained.

However, upon closer inspection of the results produced by the hidden Markov model and the neural network we noted that although some signals were impossible to detect with either method, others were much better classified with one or the other. Furthermore, the methods did not agree on every false positive, suggesting that this number could be lowered by considering both outputs. Therefore, we set out to combine the two and indeed a combination of the best scoring neural network and the best scoring Markov model yielded a correlation coefficient of 0.53, sufficiently high to be useful. The models were combined using a simple post-processing algorithm described in the Materials and methods section. Using a cut-off of 0.5 on the output from the algorithm, we obtain a sensitivity of 52%, corresponding to a maximum false positive rate of 0.1 on the test data (Figure 9). For comparison, at the 0.5 cut-off the method is thus 40.0% more sensitive than the best performing consensus pattern (L–x(2,3)–[LIVFM]–x(2,3)–L–x–[LI]), while being only 25.8% less specific. Moreover, a machine learning method has the potential to detect very divergent signals whereas the consensus pattern is completely unable to detect anything that does not conform to the pattern.

### *Validating the prediction server*

As an independent test of the predictor, we retrieved all yeast CRM1 interacting partners from the protein interaction databases MINT and DIP (44 proteins) and submitted them to

NetNES. Thirty-three of the proteins were predicted to possess a nuclear export signal and 25 of these with a score >0.7.

An example of the output of the webserver is shown in Figure 10, which is the output after submission of the amino acid sequence for the transcription factor engrailed homeobox protein. This protein, which is not in NESbase, is reported to be secreted via non-classical secretion and a requirement for this seems to be an NES allowing the protein to traverse the nuclear membrane on its way to cellular export (Prochiantz and Joliot, 2003). A, NES is reported between helix two and three in the area from 280 to 290 (Maizel *et al*., 1999) and NetNES agrees entirely with this finding.

The performance values reported in this work are based on the data set contained in NESbase, which contains most classical NES proteins such as cAMP-dependent protein kinase inhibitor (PKI), MAPK/ERK kinase 1 (MEK1) and the HIV-1 Rev protein. To evaluate the prediction server further, the literature was scanned for the most recently discovered NESs and these sequences were submitted to prediction by NetNES. Sequence homology to the training data was determined and close homologs were discarded. Where possible, the experimentally determined NES location was compared with the prediction by NetNES and the results were compiled in a table (the most recent five proteins are shown, Table III). NetNES predicts correctly the location of three of the experimentally determined NESs in these unrelated sequences.

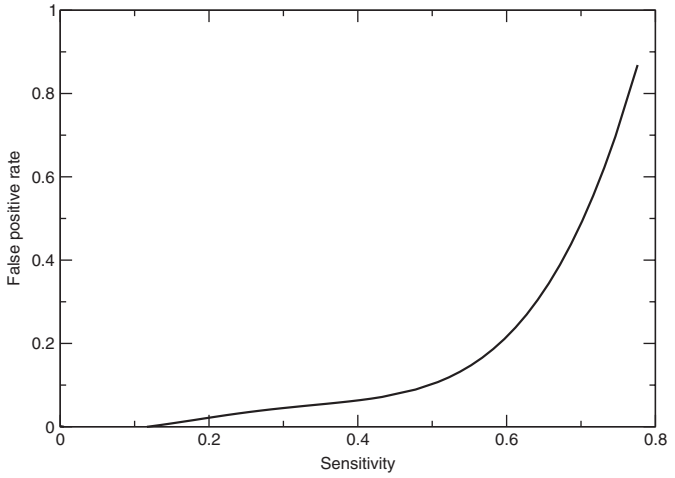


**Fig. 9.** Sensitivity and false positive rate of NetNES. This receiver operating characteristic (ROC) curve was created from test set scores using a cut-off value of 0.001 to estimate the total number of negatives. See Materials and methods for details.

## Discussion

The NES consensus as it appears from a sequence logo clearly reveals a pattern of large hydrophobic residues—primarily leucines—as the most conserved feature of NESs. With leucine being the most abundant amino acid, such a pattern is expected to be relatively frequent in proteins – and it is: The LxxLxL pattern (with L being either L, I, V, F or M) matches no less than 507 times in the data set and only 51 of these correspond to NESs. This reveals a problem: the LxxLxL pattern is abundant but does not even describe an NES satisfactorily, as only 50 (75%) of the NESs in the data set contain this pattern. In addition to the hydrophobic residues, glutamate, aspartate and serine seem to be important for NESs, since they are over-represented in certain regions near the conserved hydrophobic residues. In conclusion, an NES cannot be described by its hydrophobic pattern only. Detection of less apparent sequence features, secondary structure, flexibility and/or surface exposure is necessary for correct classification of leucine-rich regions with and without NES activity.

The secondary structure of the nine NESs for which structural information is available indicates a preference for

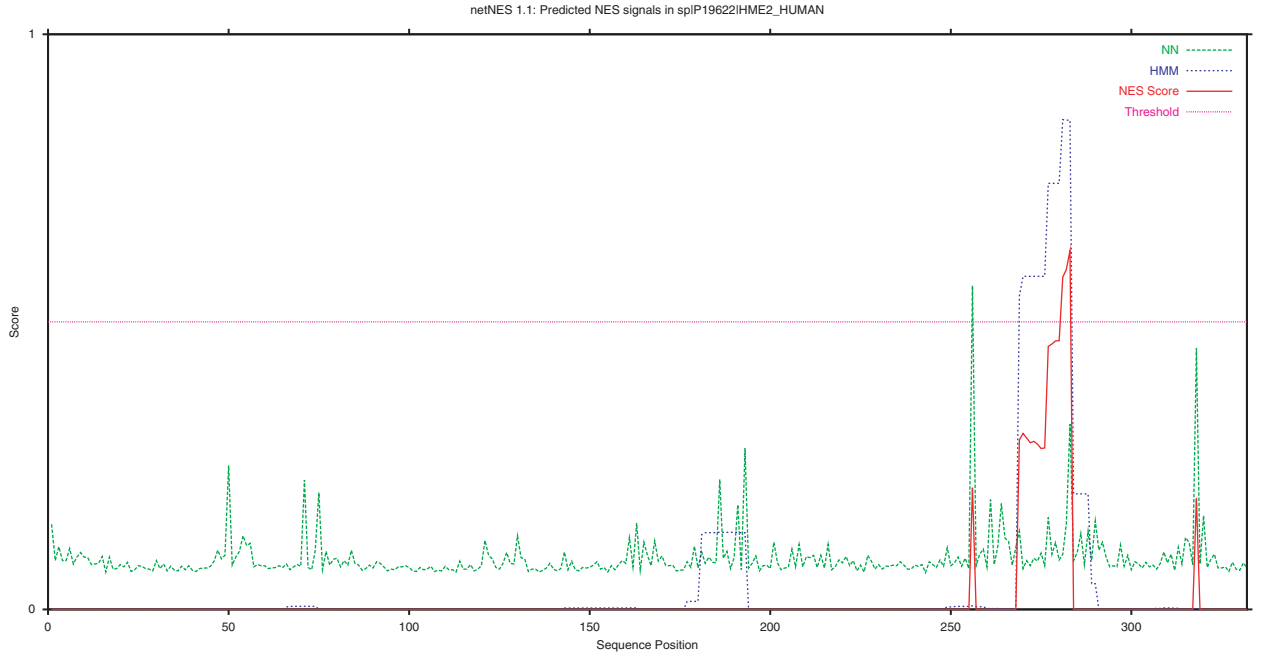


**Fig. 10.** NetNES prediction on engrailed homeobox protein. NetNES predicts a NES between residues 280 and 290, as has been found experimentally.

**Table III.** NetNES evaluation of the most recently found nuclear export signals

| Protein | Acc | Reported NES | Predicted NES | Reference |
|---|---|---|---|---|
| AID | NP 065712.1 | 183–198 | 173–183 | Ito *et al*. (2004) |
| Nipah virus V protein | NP 112023.1 | 174–192 | 3–13 | Rodriguez *et al*. (2004) |
| GRV ORF3 | NP 619660.1 | 148–156 | 142–152 | Ryabov *et al*. (2004) |
| Transducer of ERBB2 | AAH31406.1 | 2–14 | 1–10 | Maekawa *et al*. (2004) |
| Mex67 | NP 595996.1 | 434–509 | 265–275 | Thakurta *et al*. (2004) |

Sequences for which an NES was experimentally verified and its position reported but which do not have homologs in NESbase were submitted to NetNES. AID, activation-induced deaminase (*H.sapiens*); GRV ORF3, groundnut rosette virus long-distance RNA movement protein; Mex67, mRNA export factor (*S.pombe*). Transducer of ERBB2 is also from *H.sapiens*.

α-helical structure in NES regions, at least in the beginning of the signal. In fact, six of the NES structures could be structurally aligned on the first three hydrophobic residues comprising the LxxxLxxL motif, and in four of these the first hydrophobic residue was preceded by a glutamate. Neither of these features is representative of the data set since only 54% contain the LxxxLxxL motif and only 22% of these have a glutamate at the position preceding the first hydrophobic residue. Nevertheless, PSIPRED prediction of secondary structure also indicates a strong preference for α-helical structure (70.1%) and a bias against β-strand (6%) in the N-terminal end of the signal (positions P-11–P-4). The limited amount of experimental, structural data increases the risk that the data could be biased and the structural analysis should therefore be treated with some caution, although the trends are very clear.

When looking at secondary structure in a larger region around NESs, it would appear that most are located close to the border between secondary structure elements and that most of them have glutamates located on the NES helix but opposite the hydrophobic residues, as would be expected from an amphipatic helix. Amphipatic helices are found on the surface of proteins where one side of the helix faces towards the hydrophobic core of the protein and the other side is available for interaction with the aqueous environment. This facilitates detection by and binding of CRM1 and perhaps other proteins to the NES.

The fact that the large hydrophobic residues are the most conserved in an NES suggests that the interaction with CRM1 is mediated through these hydrophobic residues. This is supported by experiments providing evidence that short segments containing only the hydrophobic residues can mediate export of a reporter protein (Fischer *et al*., 1995; Wen *et al*., 1995; Stommel *et al*., 1999; Watanabe *et al*., 2000); see la Cour *et al*. (2003) for additional references.

This discovery that short NES peptides can mediate export when fused to the N-terminus of a reporter protein has certain implications. If an α-helical secondary structure is required for interaction, this structure must then be able to form independently when fused to the new protein. This need not be the case, however – it is possible that the major structural requirements for the nuclear export signal are availability and flexibility of the signal, properties we would expect from a peptide fused to a protein. We believe that this study supports the latter option. Our work indicates that the signal is usually located in highly flexible, surface-accessible regions. Specifically, the hydrophobic anchor position, which presumably is the most conserved position in the signal, is always located in a flexible and/or exposed region. Further support comes from the fact that not all NESs are located in α-helices, meaning that this structure cannot be an ultimate requirement.

Solvent exposure is, of course, a requirement for a molecular interaction to take place, α-helical structure or not. When exposure is hampered by tertiary structure, the key might just be flexibility allowing conformational changes to uncover the NES or even allow local unfolding upon interaction with CRM1. It is worth noting that serine and glutamate (and glutamine) are the most flexible surface-exposed side chains and leucine is the most flexible buried side chain (Zhao *et al*., 2001), suggesting another reason for the over-representation of these amino acids.

Several examples of NES activity being regulated by masking of the signal exist (Li *et al*., 1998; Stommel *et al*., 1999; Seimiya *et al*., 2000; Heerklotz *et al*., 2001; Kobayashi *et al*., 2001; Craig *et al*., 2002). Buried NESs can be exposed upon conformational changes or NESs can be buried upon interaction with other molecules. We suggest that the interaction between CRM1 and NES leads to local unfolding of the NES region, whereby the otherwise buried hydrophobic residues of the signal become available for interaction with the CRM1 receptor. Overall negative charge and glutamates protruding from an α-helix could help in creating the initial contact between CRM1 and the NES. Owing to its flexibility, the NES region is able to unfold upon interaction with CRM1 and thereby expose its hydrophobic residues in the correct conformation to the receptor. As this study suggests a role for the context in which NESs occur in nature, it would be of great interest to investigate experimentally the impact of glutamate, aspartate and serine mutations, as work done so far has focused only on the large hydrophobic residues.

Sequence diversity and a structure-based pattern both contribute to making NES prediction a difficult task. A possible reason for the relatively low performance of both the neural network and the hidden Markov model could be lack of available data. A neural network receiving a long amino acid segment as input has a large number of parameters and would therefore require a correspondingly large data set, which currently is not available. If more data were available, a larger sequence window might be advantageous as this would capture the structural elements that we have shown to be flanking the signal in some cases. However, we do obtain a tolerable performance from the limited data available by combining a neural network and a hidden Markov model. The performance of the predictor is sufficiently high to allow for identification of new NES-containing proteins and it is very suitable for the prediction of the location of a possible signal in the protein sequence.

Several proteins, including some in the data set used in this study, have been shown to contain more than one NES, and this could well be the case for other proteins in the data set. For lack of alternatives, we used the remainder of the sequences for negative examples of NES. Although we omitted sequences for which others have speculated on the presence of additional NES regions, we cannot exclude the possibility that some of the false positives are in fact true NESs, meaning that the actual performance of the method is higher and has increased specificity.

In summary, according to this study, features such as flexibility and/or propensity of local unfolding appear to have an impact on NES function. When reliable predictors of such properties become available, inclusion of these data could be the subject of future improvement of the NES predictor. For now, we are able to predict NESs with reasonable performance and offer the best method currently available for NES prediction.

### *Availability*

The neural network/hidden Markov model-based prediction method, NetNES, for prediction of nuclear export signals is available by following the link 'CBS prediction servers' from http://www.cbs.dtu.dk/ or at the specific URL http://www.cbs.dtu.dk/services/NetNES/.

## Acknowledgements

This work was supported by grants from the Danish National Research Foundation, the Danish Center for Scientific Computing, the Danish Natural Science Research Council, NeuroSearch A/S (to L.K.) and The John and Birthe Meyer Foundation (to T.l.C.).

## References

Appel,R.D., Bairoch,A. and Hochstrasser,D.F. (1994) *Trends Biochem. Sci.*, **19**, 258–260.
Askjaer,P., Jensen,T.H., Nilsson,J., Englmeier,L. and Kjems,J. (1998) *J. Biol. Chem.*, **273**, 33414–33422.
Baldi,P. and Brunak,S. (2001) *Bioinformatics: The Machine Learning Approach*. MIT Press, Boston.
Bernad,R., van der Velde,H., Fornerod,M. and Pickersgill,H. (2004) *Mol. Cell. Biol.*, **24**, 2373–2384.
Bischoff,F.R. and Gorlich,D. (1997) *FEBS Lett.*, **419**, 249–254.
Bischoff,F.R., Krebber,H., Smirnova,E., Dong,W. and Ponstingl,H. (1995) *EMBO J.*, **14**, 705–715.
Blom,N., Hansen,J., Blaas,D. and Brunak,S. (1996) *Protein Sci.*, **5**, 2203–2216.
Bogerd,H.P., Fridell,R.A., Benson,R.E., Hua,J. and Cullen,B.R. (1996) *Mol. Cell. Biol.*, **16**, 4207–4214.
Brunet,A., Kanai,F., Stehn,J., Xu,J., Sarbassova,D., Frangioni,J.V., Dalal,S.N., DeCaprio,J.A., Greenberg,M.E. and Yaffe,M.B. (2002) *J. Cell Biol.*, **156**, 817–828.
Brzovic,P.S., Rajagopal,P., Hoyt,D.W., King,M.C. and Klevit,R.E. (2001) *Nat. Struct. Biol.*, **8**, 833–837.
Chen,X., Vinkemeier,U., Zhao,Y., Jeruzalmi,D., Darnell,J.E. and Kuriyan,J. (1998) *Cell*, **93**, 827–839.
Clore,G.M., Omichinski,J.G., Sakaguchi,K., Zambrano,N., Sakamoto,H., Appella,E. and Gronenborn,A.M. (1994) *Science*, **265**, 386–391.
Cokol,M., Nair,R. and Rost,B. (2000) *EMBO Rep.*, **1**, 411–415.
Craig,E., Zhang,Z.K., Davies,K.P. and Kalpana,G.V. (2002) *EMBO J.*, **21**, 31–42.
Durbin,R.M., Eddy,S.R., Krogh,A. and Mitchison,G. (1998) *Biological Sequence Analysis*. Cambridge University Press, Cambridge.
Engel,K., Kotlyarov,A. and Gaestel,M. (1998) *EMBO J.*, **17**, 3363–3371.
Englmeier,L., Fornerod,M., Bischoff,F.R., Petosa,C., Mattaj,I.W. and Kutay,U. (2001) *EMBO Rep.*, **2**, 926–932.
Essen,L.O., Perisic,O., Katan,M., Wu,Y., Roberts,M.F. and Williams,R.L. (1997) *Biochemistry*, **36**, 1704–1718.
Fabbro,M. and Henderson,B.R. (2003) *Exp. Cell Res.*, **282**, 59–69.
Fischer,U., Huber,J., Boelens,W.C., Mattaj,I.W. and Luhrmann,R. (1995) *Cell*, **82**, 475–483.
Fornerod,M., Ohno,M., Yoshida,M. and Mattaj,I.W. (1997) *Cell*, **90**, 1051–1060.
Fukuda,M., Asano,S., Nakamura,T., Adachi,M., Yoshida,M., Yanagida,M. and Nishida,E. (1997) *Nature*, **390**, 308–311.
Gama-Carvalho,M. and Carmo-Fonseca,M. (2001) *FEBS Lett.*, **498**, 157–163.
Haasen,D., Kohler,C., Neuhaus,G. and Merkle,T. (1999) *Plant J.*, **20**, 695–705.
Heerklotz,D., Doring,P., Bonzelius,F., Winkelhaus,S. and Nover,L. (2001) *Mol. Cell. Biol.*, **21**, 1759–1768.
Hobohm,U., Scharf,M., Schneider,R.and Sander,C. (1992) *Protein Sci.*, **1**, 409–417.
Hope,T.J. (1999) *Arch. Biochem. Biophys.*, **365**, 186–191.
Ito,S., Nagaoka,H., Shinkura,R., Begum,N., Muramatsu,M., Nakata,M. and Honjo,T. (2004) *Proc. Natl Acad. Sci. USA*, **101**, 1975–1980.
Jones,D.T. (1999) *J. Mol. Biol.*, **292**, 195–202.
Jones,T.A., Zou,J.Y., Cowan,S.W. and Kjeldgaard,M. (1991) *Acta Crystallogr.*, **A47** (Pt 2), 110–119.
Kau,T.R. and Silver,P.A. (2003) *Drug Discov. Today*, **8**, 78–85.
Kau,T.R., Way,J.C. and Silver,P.A. (2004) *Nat. Rev. Cancer*, **4**, 106–117.
Kobayashi,T., Kamitani,W., Zhang,G., Watanabe,M., Tomonaga,K. and Ikuta,K. (2001) *J. Virol.*, **75**, 3404–3412.
Kraulis,P.J. (1991) *J. Appl. Crystallogr.*, **24**, 946–950.
Kudo,N., Matsumori,N., Taoka,H., Fujiwara,D., Schreiner,E.P., Wolff,B., Yoshida,M.and Horinouchi,S. (1999a) *Proc. Natl Acad. Sci. USA*, **96**, 9112–9117.
Kudo,N., Taoka,H., Toda,T., Yoshida,M. and Horinouchi,S. (1999b) *J. Biol. Chem.*, **274**, 15151–15158.
Kuge,S., Arita,M., Murayama,A., Maeta,K., Izawa,S., Inoue,Y. and Nomoto,A. (2001) *Mol. Cell. Biol.*, **21**, 6139–6150.
la Cour,T., Gupta,R., Rapacki,K., Skriver,K., Poulsen,F.M. and Brunak,S. (2003) *Nucleic Acids Res*, **31**, 393–396.
Li,Y., Yamakita,Y. and Krug,R.M. (1998) *Proc. Natl Acad. Sci. USA*, **95**, 4864–4869.
Lindsay,M.E., Holaska J.M., Welch K., Paschal B.M. and Macara I.G. (2001) *J. Cell Biol.*, **153**, 1391–1402.
Maekawa,M., Yamamoto,T. and Nishida,E. (2004) *Exp. Cell Res.*, **295**, 59–65.
Maizel,A., Bensaude,O., Prochiantz,A. and Joliot,A. (1999) *Development*, **126**, 3183–3190.
Matthews,B.W. (1975) *Biochim. Biophys. Acta*, **405**, 442–451.
McKinsey,T.A., Zhang,C.L. and Olson,E.N. (2001) *Mol. Cell. Biol.*, **21**, 6312–6321.
Meng,W., Swenson,L.L., Fitzgibbon,M.J., Hayakawa,K., Ter Haar,E., Behrens,A.E., Fulghum,J.R. and Lippke,J.A. (2002) *J. Biol. Chem.*, **277**, 37401–37405.
Merrit,E.A. and Bacon,D.J. (1997) *Methods Enzymol.*, **277**, 505–524.
Nair,R., Carter,P. and Rost,B. (2003) *Nucleic Acids Res.*, **31**, 397–399.
Neville,M. and Rosbash,M. (1999) *EMBO J.*, **18**, 3746–3756.
Neville,M., Stutz,F., Lee,L., Davis,L.I. and Rosbash,M. (1997) *Curr. Biol.*, **7**, 767–775.
Nielsen,H., Engelbrecht,J., Brunak,S. and von Heijne,G. (1997) *Protein Eng.*, **10**, 1–6.
Ohno,M., Segref,A., Bachi,A., Wilm,M. and Mattaj,I.W. (2000) *Cell*, **101**, 187–198.
Ossareh-Nazari,B. and Dargemont,C. (1999) *Exp. Cell Res*, **252**, 236–241.
Ossareh-Nazari,B., Bachelerie,F. and Dargemont,C. (1997) *Science*, **278**, 141–144.
Ossareh-Nazari,B., Gwizdek,C. and Dargemont,C. (2001) *Traffic*, **2**, 684–689.
Prochiantz,A. and Joliot,A. (2003) *Nat. Rev. Mol. Cell Biol.*, **4**, 814–819.
Qin,B.Y., Chacko,B.M., Lam,S.S., de Caestecker,M.P., Correia,J.J. and Lin,K. (2001) *Mol. Cell*, **8**, 1303–1312.
Rodriguez,J.J., Cruz,C.D. and Horvath,C.M. (2004) *J. Virol.*, **78**, 5358–5367.
Ryabov,E.V., Kim,S.H. and Taliansky,M. (2004) *J. Gen. Virol.*, **85**, 1329–1333.
Ryan,K.J. and Wente,S.R. (2000) *Curr. Opin. Cell Biol.*, **12**, 361–371.
Schneider,T.D. and Stephens,R.M. (1990) *Nucleic Acids Res.*, **18**, 6097–6100.
Seimiya,H., Sawada,H., Muramatsu,Y., Shimizu,M., Ohko,K., Yamane,K. and Tsuruo,T. (2000) *EMBO J.*, **19**, 2652–2661.
Shannon,C.E. (1948) *Bell Syst. Tech. J.*, **27**, 379–423/623–656.
Stade,K., Ford,C.S., Guthrie,C. and Weis,K. (1997) *Cell*, **90**, 1041–1050.
Stoffler,D., Fahrenkrog,B. and Aebi,U. (1999) *Curr. Opin. Cell Biol.*, **11**, 391–401.
Stommel,J.M., Marchenko,N.D., Jimenez,G.S., Moll,U.M., Hope,T.J. and Wahl,G.M. (1999) *EMBO J.*, **18**, 1660–1672.
Strom,A.C. and Weis,K. (2001) *Genome Biol.*, **2**, REVIEWS3008.
Taura,T., Krebber,H. and Silver,P.A. (1998) *Proc. Natl Acad. Sci. USA*, **95**, 7427–7432.
Thakurta,A.G., Gopal,G., Yoon,J.H., Saha,T. and Dhar,R. (2004) *J. Biol. Chem.*, **279**, 17434–17442.

Tickenbrock,L., Cramer,J., Vetter,I.R. and Muller,O. (2002) *J. Biol. Chem.*, **277**, 32332–32338.
Vorobiev,S., Strokopytov,B., Drubin,D.G., Frieden,C., Ono,S., Condeelis,J., Rubenstein,P.A. and Almo,S.C. (2003) *Proc. Natl Acad. Sci. USA*, **100**, 5760–5765.
Wada,A., Fukuda,M., Mishima,M. and Nishida,E. (1998) *EMBO J.*, **17**, 1635–1641.
Watanabe,M., Masuyama,N., Fukuda,M. and Nishida,E. (2000) *EMBO Rep.*, **1**, 176–182.
Wen,W., Meinkoth,J.L., Tsien,R.Y. and Taylor,S.S. (1995) *Cell*, **82**, 463–473.
Yan,C., Lee,L.H. and Davis,L.I. (1998) *EMBO J.*, **17**, 7416–7429.
Zhang,Y. and Xiong,Y. (2001) *Science*, **292**, 1910–1915.
Zhao,S., Goodsell,D.S. and Olson,A.J. (2001) *Proteins*, **43**, 271–279.

**Received June 15, 2004; accepted July 28, 2004**

**Edited by P.Balaram**