7388–7399 Nucleic Acids Research, 2010, Vol. 38, No. 21 Published online 26 July 2010
doi:10.1093/nar/gkq653

Characterization and prediction of protein nucleolar localization sequences

Michelle S. Scott[1,*], François-Michel Boisvert[2], Mark D. McDowall[1], Angus I. Lamond[2] and Geoffrey J. Barton[1]

[1]Division of Biological Chemistry and Drug Discovery and [2]Wellcome Trust Centre for Gene Regulation and Expression, College of Life Sciences, University of Dundee, Dow Street, Dundee DD1 5EH, UK

Received May 31, 2010; Revised July 7, 2010; Accepted July 8, 2010

ABSTRACT

Although the nucleolar localization of proteins is often believed to be mediated primarily by non-specific retention to core nucleolar components, many examples of short nucleolar targeting sequences have been reported in recent years. In this article, 46 human nucleolar localization sequences (NoLSs) were collated from the literature and subjected to statistical analysis. Of the residues in these NoLSs 48% are basic, whereas 99% of the residues are predicted to be solvent-accessible with 42% in α-helix and 57% in coil. The sequence and predicted protein secondary structure of the 46 NoLSs were used to train an artificial neural network to identify NoLSs. At a true positive rate of 54%, the predictor's overall false positive rate (FPR) is estimated to be 1.52%, which can be broken down to FPRs of 0.26% for randomly chosen cytoplasmic sequences, 0.80% for randomly chosen nucleoplasmic sequences and 12% for nuclear localization signals. The predictor was used to predict NoLSs in the complete human proteome and 10 of the highest scoring previously unknown NoLSs were experimentally confirmed. NoLSs are a prevalent type of targeting motif that is distinct from nuclear localization signals and that can be computationally predicted.

INTRODUCTION

The nucleolus is a prominent non-membrane-contained nuclear structure known primarily as the site of ribosome biogenesis and assembly (1). In the past two decades however, the nucleolus has been shown to be involved in various other cellular functions including assembly of diverse ribonucleoprotein particles (RNPs), cell-cycle progression and proliferation regulation, as well as the response to numerous forms of cellular stress (2–6). Many of the processes that occur, at least in part, in the nucleolus require the re-location, often cyclical or conditional, of nucleoplasmic and even cytoplasmic proteins to the nucleolus (2–4,7). Consistent with this, the nucleolar proteome is large with currently over 4500 distinct human proteins that have been identified in purified nucleoli (8) and has been shown to respond dynamically to various treatments (9,10). The nucleolus thus accommodates a large and dynamic volume of cellular traffic, which presumably requires tight regulation of its protein targeting mechanisms. However, as highlighted in two recent reviews, widely accepted mechanisms of protein targeting to the nucleolus remain elusive (6,11).

In contrast, protein targeting to membrane-bound cellular compartments is well characterized and a small number of short targeting sequence motifs are predominantly used. These short targeting motifs are generally recognized by the import machinery of the target compartment. Such is the case for nuclear localization signals (NLSs) for targeting across the nuclear envelope (12), signal peptides for co-translational entry into the secretory pathway at the endoplasmic reticulum (13) as well as mitochondrial targeting peptides (14) and peroxisomal targeting signals (15). Protein localization in the nucleolus, on the contrary, is not generally well understood and is widely believed to be the result of interaction by high affinity binding to nucleolar core components such as ribosomal DNA, RNA or major protein components (16). Thus, nucleolar localization would result from retention in the nucleolus rather than targeting to this compartment.

However, in the past 15 years, numerous reports of unrelated human proteins harbouring nucleolar localization sequences (NoLSs) have been published (summarized in Table 1). Not all these motifs have been rigorously tested, but many have been shown to be sufficient for targeting reporter proteins to the nucleolus. While some of these NoLSs have been manually aligned with previously

*To whom correspondence should be addressed. Tel: +44 1382 386097; Fax: +44 1382 345893; Email: michelle@compbio.dundee.ac.uk

© The Author(s) 2010. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.5), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from http://nar.oxfordjournals.org/ at Universite de Sherbrooke on November 16, 2012

Table 1. Experimentally Validated NoLSs (EVN) dataset

Accession	Protein name	NoLS	Targets reporter protein to nucleolus[a]
NP_001012270	BIRC5	MQRKPTIRRKNLRLRRK	GFP
NP_006161	NOP2	SKRLSSRARKRAAKRRLG	β-Gal (but requires additional NLS)
NP_005336	HSPA1A	FKRKHKKDISQNKRAVRR	GFP
NP_937862	ING1b (NoLS-1)	DKPNSKRSRRQRNNENR	GFP
NP_937862	ING1b (NoLS-2)	TPKEKKAKTSKKKKRSKAKA	GFP
NP_005238	FGF3	GKGVQPRRRRQKQSPDNLEP	N/A
NP_006618	POP4	RHKRKEKKKKAKGLSARQRRELR	GFP
NP_945316	PTHLH	GKKKKGKPGKRREQEKKKRRT	β-gal
NP_003778	NOL4	KEKIQAIIDSCRRQFPEYQERAR	N/A
NP_001002	RPS7	RRILPKPTRKSRTKNKQKRPR	N/A
NP_001034800	DEDD	LKRRRA	N/A
NP_001091059	RPP38	KIKKLIPNPNKIRKPPKSKKATPK	GFP
NP_478102	CDKN2A	QLRRPRHSHPTRARRCP	GFP
NP_003133	SSB	QESLNKWKSKGRRFKGKGKGNKAAQPGSGKGK	PTB-GFP
NP_005560	LIMK2	KKRTLRKNDRKKR	GFP
NP_001997	FGF2	RSRKYTSWYVALKR	GFP
NP_477352	PI4KA	SKKTNRGSQLHKYYMKRRTL	Soybean trypsin inhibitor
NP_002383	MDM2	KKLKKRNK	Thioredoxin
NP_003945	MAP3K14	RKKRKKK	GFP
NP_078908	SAP30L	RRYKRHYK	N/A
NP_951038	MDFIC	GRCRRLANFPGRKRRRRRR	GFP
NP_848927	MTDH (NoLS-1)	KSKKKKKKKKKQGE	GFP
NP_848927	MTDH (NoLS-2)	KQIKKKKKARRET	GFP
NP_078805	CDC73 (NoLS-1)	RRAATENIPVVRRPDRK	GFP
NP_078805	CDC73 (NoLS-2)	KKKQGCQRENETLIQRRK	GFP
NP_078905	MLF1IP	MAPRGRRRPRPHRSEGARRSKNTLERTHS	GFP
NP_060239	G2E3	RKHDDCPNKYGEKKTKEK	N/A
NP_077289	NOL12	KRKHPRRAQDSKKPPRAPRTSKAQRRR	GFP fused to rat NOL12-NoLS
NP_039252	NRG1	MSERKEGRGKGKGKKKERGSGKK	GFP
NP_055318	UTP20	KKKMKKHKNKSEAKKRK	GFP
NP_849193	STT3B	KQKYLSKKTTKRKRGYIKNKLVFKKGKKISKKTV	GFP
NP_068810	RELA	EQPKQRGMRFRYKCEGRSAGSIPGER	N/A
NP_112578	INO80B	HGHGVHKKKHKKHKKKHKKKHH	N/A
AAB60345	L1 ORF2	RLKIKGQRKIYQANGKQKK	N/A
AAH01024	GNL3	KRPKLKKASKRMTCHKRYKIQKKVREHHRKLRLEAKKQGHKKPRK	N/A
NP_002511	NPM1	QDLWQWRKSL	GFP
NP_937983	TERT	MPRAPRCRAVRSLLR	GFP
NP_003277	TOP1	NKKKKPKKE	N/A
NP_796375	MIDN	QQKRLRRKARRDARGPYHWSPSRKAGRS	GFP
NP_004851	FXR2	RPQRRNRSRRRRNR	N/A
NP_000347	TCOF1	KRKKDKEKKEKKKKAKKASTKDSESPSQKKKKKKKKTAEQTV	GFP
NP_004695	RRP9	GQEHRLGRWWRIKEARNSVCIIPLRRVPVPPAAGS	N/A
NP_150241	PML	DRPLVFFDLKIDN	GFP
NP_061940	GNL3L	MMKLRHKNKKPGEGSKGHKKISWPYPQPA KQNGKKATSKVPSAPHFVHPN	GFP
NP_004251	RECQL4	KQAWKQKWRKK	GFP
NP_068778	PPP1R11	HRKGRRR	N/A

[a]Indicates whether this NoLS has been shown to target a reporter protein to the nucleolus when fused to it. The reporter protein chosen is indicated and references are provided in Supplementary File 1.

known NoLSs, no systematic study of these motifs has been reported. Here, we investigate the characteristics of these experimentally validated NoLSs and use them as a training set to computationally predict NoLSs in the entire human proteome.

MATERIALS AND METHODS

Datasets

Positive examples of NoLSs were manually curated from the literature and are referred to as the experimentally validated NoLSs (EVN, listed in Table 1 and detailed in Supplementary File 1) set.

Three types of negatives were considered:

- Non-NoLS NLSs that were manually curated from the literature and the NLSdb (17) and are listed in Supplementary File 2.
- Randomly chosen sequences of length 20 from cytoplasmic non-nucleolar proteins as annotated by Uniprot (18).
- Randomly chosen sequences of length 20 from nucleoplasmic non-nucleolar proteins as annotated by Uniprot (18).

Downloaded from http://nar.oxfordjournals.org/ at Universite de Sherbrooke on November 16, 2012

The training/testing dataset should be a representative set that maximizes coverage while minimizing redundancy (19,20). Redundancy filtering was performed by ensuring that all the corresponding full-length proteins from which the sub-sequences are extracted to generate the datasets are <30% identical over their entire sequence to any other corresponding full-length protein used to generate the dataset. In addition to this, we also verified that our datasets are non-redundant by extending all the sub-sequences considered to a size of 50 (the length of the longest EVN NoLS) and aligning them pairwise using the fasta program (version 35.04) (21). All extended NoLS pairs have at most 13 exact matches in local alignments, representing <30% sequence identity between the pairs.

For the purpose of training the ANN, several different combinations of the datasets were investigated and their performance compared by cross-validation. The one that was settled on consists of unbalanced datasets comprising 20 copies of the positive examples, 5 copies of the non-NoLS NLSs negatives, ~1000 cytoplasmic negatives and 180 nucleoplasmic negatives. When 3-fold cross-validation was performed, care was taken to ensure that all copies of a given sequence (for NoLSs and non-NoLS NLSs which were used in more than one copy) were placed in the same group.

Encoding

For the sequence encoding, windows of 13 residues in size were sparsely encoded in a binary manner using a reduced alphabet of size 12 with the follow groupings: {K, R, Q, P, H, ED, STY, N, C, W, ILVAMG, F}. For example, the sequence NSAT would be encoded as the binary vector 0000000100000000001000000000000000010000000100000. This reduced alphabet was chosen to ensure that frequent residues in NoLSs are represented as singlets while under-represented residues in NoLSs are grouped by chemical similarity. Other sequence encodings were considered but did not outperform the encoding described here as assessed by cross-validation.

For the sequence encoding, a window size of 13 was chosen for several reasons: (i) bipartite NLSs are between 15 and 17 residues in length according to Prosite (22) and thus a window size shorter than this might minimize the number of NLSs wrongly predicted as NoLSs, (ii) larger window sizes lead to larger artificial neural networks (ANNs) and a higher possibility to overfitting, (iii) the accuracy by 3-fold cross-validation is substantially worse when the window size is greater than 16 or smaller than 11, and 4) an odd number for the window size makes it easier to assign a score to the middle residue.

Additional information including protein characteristics and secondary structure were also considered and encoded using nine floating point numbers:

- a representation S_L of the length L of the protein

  $S_L = 1$ if $L > 400$

  otherwise, $S_L = 1 - \frac{400 - L}{400}$

  400 was chosen as a threshold as this is the approximate average length of human proteins as defined by IPI version 3.40 (23).

- a representation D of the relative distance between the sub-sequence considered and the middle of the full-length protein

  $$D = \frac{|x - m|}{m}$$

  where x is the position of the subsequence considered and m is the position of the middle of the protein.

- and 7 measures of protein secondary structure all predicted by Jpred (24) over a region R covering the window of size 13 considered and three flanking residues on either side:
  - the proportion of residues in R predicted as belonging to an α-helix
  - the proportion of residues in R predicted as belonging to a β-sheet
  - the proportion of residues in R predicted as located in a coil
  - the average confidence of the three above predictions over region R, as estimated by Jpred (24)
  - the proportion of buried residues in R predicted at a relative solvent accessibility threshold of >25%
  - the proportion of buried residues in R predicted at a relative solvent accessibility threshold of >5%
  - the proportion of buried residues in R predicted at a relative solvent accessibility threshold of >0%

When only the sequence information is used, a binary vector of size 156 is created (window of size 13 × alphabet of size 12). If in addition to sequence, protein characteristics and secondary structure are considered, a vector of size 165 (156 + 9) is created.

ANNs

The Stuttgart Neural Network Simulator (SNNS; http://www.ra.cs.uni-tuebingen.de/SNNS/) was used to train ANNs for the purpose of predicting NoLSs. Many different combinations of neural network architecture and parameters were investigated. Most performed equally well, indicating that the method is relatively insensitive to parameter changes, and many of the default settings were chosen. The combination settled on is described here. ANNs were built with either 156 or 165 input nodes (depending on the encoding used, see 'Encoding' section), 9 hidden nodes and 1 output node. The chosen target outputs were 0 for non-NoLSs and 1 for NoLSs. The learning function used was batch backpropopagation, the initialization function was Randomize_Weights and the update function was Topological_Order.

During 3-fold cross-validation, ANNs were trained until the prediction performance on the validation set started decreasing (~4000 cycles).

For the receiver operating characteristic (ROC) plots, the ANN was trained and validated on all three types of negatives combined and it is just for testing purposes that

Downloaded from http://nar.oxfordjournals.org/ at Universite de Sherbrooke on November 16, 2012

the three types of negatives were considered separately as well as combined (see Figure 3).

Characterization of predicted NoLS-containing proteins

For the characterization of predicted NoLS-containing proteins, 'experimental' subcellular localization annotations were downloaded from Uniprot (18) for all human proteins. DAVID (25) was used to compare the GO biological process term enrichment between the list of predicted NoLS-containing proteins that exist in RefSeq and the list of all human RefSeq proteins that were considered by our predictor as background.

Cell culture and transfection

The human osteosarcoma cell line U2OS was cultured as adherent cells in Dulbeccos's modified eagle medium (DMEM) (Invitrogen) supplemented with 10% fetal bovine serum, 100 U/ml penicillin/streptomycin and 2 mM L-glutamine. Transfection was done using Effectene (QIAGEN) as per the manufacturer protocol.

Cloning

The oligonucleotides corresponding to each NoLS considered (see Supplementary File 3 for their nucleotide sequences and Table 4 for their amino acid sequences) were annealed by first heating them at 95°C and then letting them cool down to room temperature. The resulting double-stranded DNA was then cloned into pEGFP-C1 (Clontech) using the restriction enzymes Bgl II and Kpn I.

Immunofluorescence

Cells were grown on glass coverslips and fixed with 1% paraformaldehyde in PBS for 10 min. Cells were then permeabilized in PBS containing 0.5% Triton X-100 for 10 min and mounted on slides with Vectashield (Vector Laboratories Inc.) containing DAPI. Fluorescence imaging was performed on a DeltaVision Spectris widefield deconvolution microscope (Applied Precision), using a CoolMax charge-coupled device camera (Roper Scientific). Cells were imaged using a $60 \times$ NA 1.4 Plan-Apochromat objective (Olympus) and the appropriate filter sets (Chroma Technology Corp.), with 20 optical sections of 0.5 μM each acquired. SoftWorX software (Applied Precision) was used for both acquisition and deconvolution.

RESULTS

General NoLS characteristics

A dataset of experimentally validated NoLSs was assembled by extensive manual curation of the literature. Reported NoLSs of length >50 residues were discarded as their critical residues have likely not been precisely defined and/or the NoLS might form a signal patch and exist only in the folded protein. The remaining 46 NoLSs are shown in Table 1. These will be referred to as the experimentally validated NoLS (EVN) set.

Visual inspection of the EVN sequences reveals a high proportion of basic amino acids. In fact, 48% of the residues found in these sequences are lysines or arginines. The average residue frequency for all amino acids in EVN sequences is shown in Supplementary File 4.

The secondary structure predictor Jpred 3 (24) was used to analyze the protein regions that contain NoLSs (Figure 1). EVN sequences are localized in regions predicted to be almost uniquely α-helices or coils (Figure 1A) and found predominantly at the surface of proteins (Figure 1B). An analysis of the position of experimentally validated NoLSs in full-length proteins shows that known NoLSs localize predominantly at the ends of proteins (Figure 1C). In fact, 22 of the 46 NoLSs examined are found in the 25% of residues closest to the protein termini. NoLSs are thus localized in protein regions that are easily accessible.

NoLS vs NLS

NLSs target proteins to the nucleus. Numerous and diverse NLSs have been reported and mechanisms of recognition of NLSs have been extensively studied (12,26). NoLSs and NLSs have very similar amino acid compositions (a high prevalence of basic residues in both cases) and while there is mounting evidence that these two types of signals are recognized as different by the cell, little attention has been given to distinguishing and systematically characterizing both types of signals. NoLSs and NLSs can be collectively grouped into three classes:

- NLS-only signals that target proteins to the nucleus but do not cause significant accumulation in the nucleolus [e.g. PTMA is nucleoplasmic and harbours a bipartite and non-NoLS NLS (27)].
- NoLS-only signals that cause proteins to accumulate in the nucleolus but are unable to mediate nuclear envelop translocation. These are usually found in proteins that also contain an NLS-only signal. For example, the proteins NOP2 (28) and PPP1R11 described below.
- Joint NoLS-NLS regions which can both target proteins across the nuclear envelope and cause proteins to accumulate in the nucleolus. For example, UTP20 is reported to contain overlapping NLS and NoLS near its C-terminus (29).

To confirm that these signals are necessary and sufficient for this targeting, they are usually fused to reporter proteins and visualized by microscopy (see Table 1 for examples of experimentally confirmed NoLSs).

Several proteins are reported to contain two 'NLSs', one of which seems to allow entry into the nucleus (an NLS-only signal) and the other which targets nuclear proteins to the nucleolus (an NoLS-only signal). For example, PPP1R11 (protein phosphatase-1 inhibitor-3) is mainly nucleolar. It has two basic stretches that have different targeting roles. The most N-terminal basic motif (residues 32–37) serves as an NLS and the protein accumulates in the cytoplasm when this signal is mutated. In contrast, a C-terminal motif (residues 94–100) functions as an NoLS and the protein is nuclear but non-nucleolar when this motif is absent (30).

Downloaded from http://nar.oxfordjournals.org/ at Universite de Sherbrooke on November 16, 2012

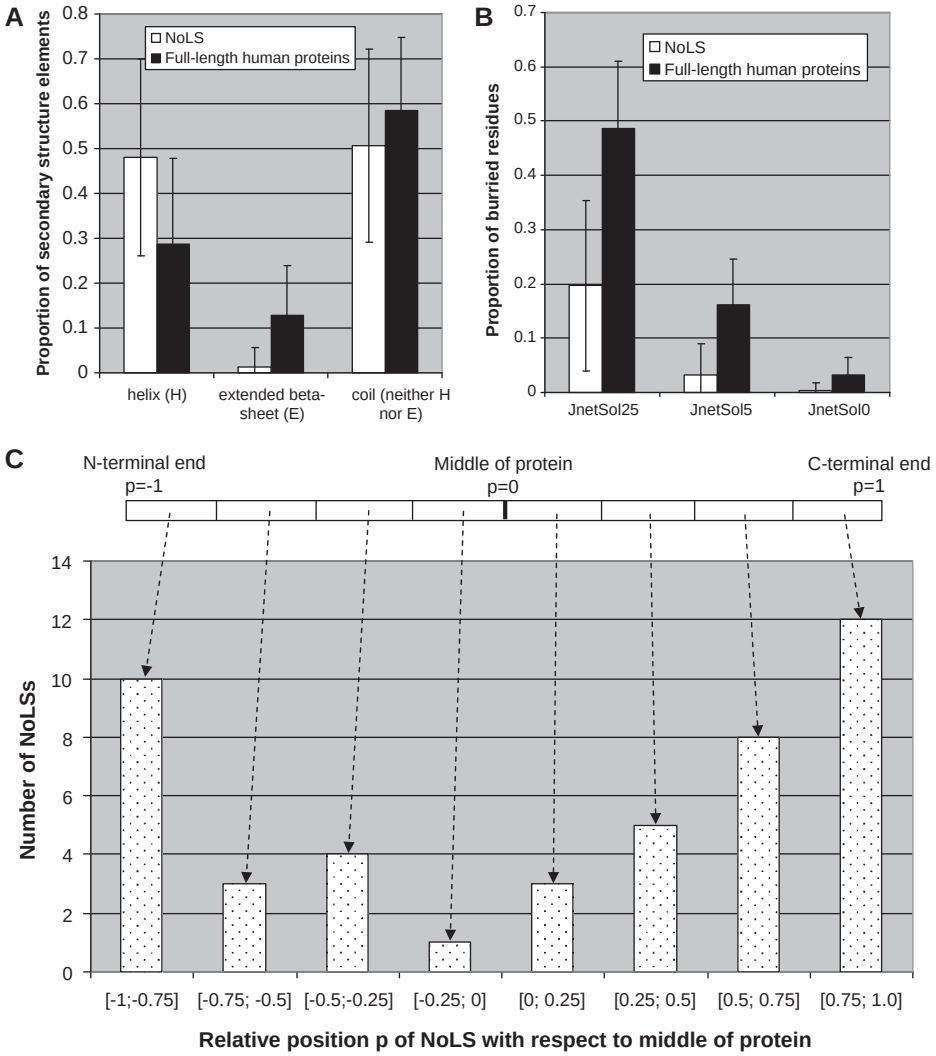

Figure 1. NoLS characteristics. (**A**) NoLSs are predominantly found in regions predicted by Jpred (24) as α-helices or coils and very rarely in regions predicted as extended β-strands. (**B**) NoLSs localize predominantly at the surface of proteins as predicted by Jnet (24) either at relative solvent accessibility thresholds <25% (JnetSol25), <5% (JnetSol5) or 0% (JnetSol0). (**C**) NoLSs are found predominantly at the ends of proteins. The error bars represent standard deviation.

Prediction of NoLSs using ANNs

The EVN dataset was used to investigate whether known NoLSs can be identified computationally and predicted at the proteome level. ANNs were chosen as a machine learning method to predict NoLSs because they perform well at pattern recognition tasks and have been used successfully to identify other protein targeting motifs (31,32). For this task, the aim is to differentiate between NoLS and non-NoLS sequences. For training purposes, the ANN thus requires both positive examples of NoLSs (the EVN dataset) and examples of sequences that do not target proteins to the nucleolus (referred to as the negative training set). As described in the 'Materials and methods' section, the negative training set was generated by combining three groups of non-NoLS sequences: (i) randomly chosen protein sub-sequences of 20 residues from cytoplasmic proteins not annotated as localizing to the nucleolus, (ii) randomly chosen protein sub-sequences of 20 residues from nucleoplasmic proteins not annotated as localizing to the nucleolus and (iii) reported NLSs for which there is no evidence that they also localize proteins to the nucleolus (NLS-only signals, as described above). As NLSs and NoLSs have similar amino acid compositions, NLSs represent the most difficult group of negatives to predict against. Non-NoLS NLSs used in the negative training set were identified by manual curation of the literature and of NLSdb (17). However, in assembling this dataset, it became obvious that many reported NLSs might also be NoLSs (joint NoLS-NLS regions as described above) or are found in nucleolar proteins and no investigation has been performed to check whether these NLSs are also NoLSs. For example, NLS27 and NLS30 from NLSdb (17) refer to the NLS of the protein LEF1 described in (33). However, while some microscopy pictures in (33) show LEF1 accumulating in structures that resemble nucleoli, and Entrez Gene annotates LEF1 as being nucleolar, no further investigation has been undertaken to clarify the true nature of the LEF1

Downloaded from http://nar.oxfordjournals.org/ at Universite de Sherbrooke on November 16, 2012

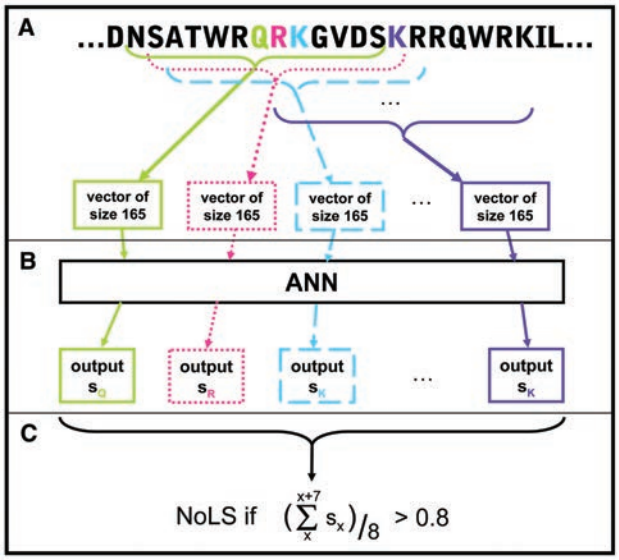

Figure 2. Prediction of NoLSs using an ANN. (**A**) Sequence windows of size 13 overlapping with an offset of 1 are sparsely encoded into binary vectors of size 165 based on their amino acid sequence, position within the full-length protein sequence and elements of secondary structure. (**B**) The encoded vectors are fed to the ANN which outputs one score for each input window, attributed to the central residue of the window. (**C**) Peptides of length 20 are predicted as NoLSs if the average score of the 8 windows of size 13 they contain is >0.8.

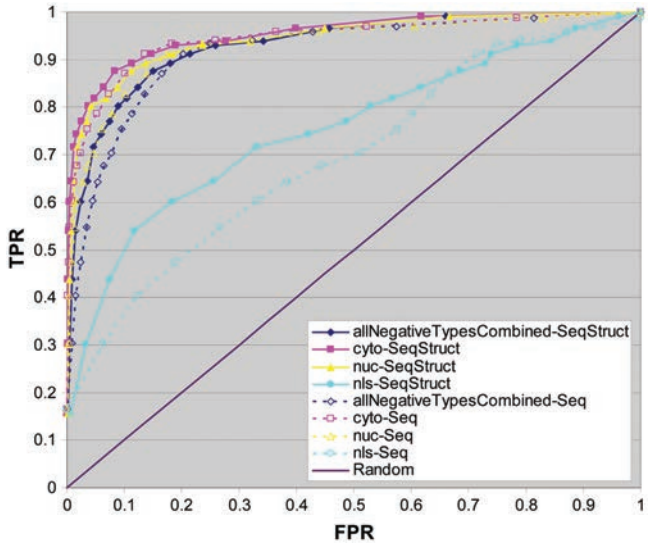

Figure 3. ROC plots. The predictor was trained by 3-fold cross-validation using all types of negatives combined. The true positive rates (TPRs) versus false positive rates (FPRs) are plotted for the three different types of negatives tested collectively (allNegativeTypesCombined) and separately: randomly chosen cytoplasmic sequences (referred to as cyto), randomly chosen nucleoplasmic sequences (referred to as nuc) and curated non-NoLS NLSs (labelled nls). The accuracy measures of two encodings are shown: encodings based only on sequence (Seq) and encodings based on both sequence and additional structure elements (Seq-Struct). The diagonal line indicates the performance that would be expected at random.

'NLS'. Reported NLSs found in proteins localized to the nucleolus were excluded from the negative training set.

Positive and negative training set sequences were encoded as described in Figure 2 and the 'Materials and methods' section. ANNs were built using the SNNS (http://www.ra.cs.uni-tuebingen.de/SNNS/).

Measures of accuracy

Cross-validation. Three-fold cross-validation experiments were performed to measure the accuracy of the predictor. The positive and negative datasets were randomly divided into three non-overlapping sets used respectively for training, validating and testing the ANN. The reported accuracy is the average of the different training, validating and testing combinations. Figure 3 summarizes the performance of the predictor as a ROC plot in which the true positive rate (TPR) is plotted against the false positive rate (FPR) of the predictor. The predictor was trained on the combination of all three types of negative examples as described above and subsequently tested on this combination of negatives (points labelled allNegativeTypesCombined). To investigate how well the predictor performs on the different types of negatives, in Figure 3 we also provide a breakdown of their estimated accuracy separately. This was done by training the predictor in a cross-validation manner on all three types of negatives combined and then considering each of these types of negatives separately for testing. As shown in Figure 3, including secondary structure information as well as sequence (solid lines) consistently results in higher accuracy compared to using only sequence (dashed lines) for all negative types. As expected, the predictor performs better on negatives randomly generated from nucleoplasmic or cytoplasmic non-nucleolar proteins than when tested with reported NLSs. To yield low FPRs while maintaining a reasonably high TPR, the threshold to predict NoLSs was set to an average output score of 0.8 over 8 consecutive windows (as described in the 'Materials and methods' section and in Figure 2). At this score, the average TPR is measured to be 54% and the FPRs are measured to be 0.26% for the randomly chosen cytoplasmic sequences, 0.80% for the randomly chosen nucleoplasmic sequences and 12% for the NLSs.

Independent validation on NoLS-containing proteins of human-infecting viruses. Numerous and diverse viral proteins have been shown to localize in the nucleoli of their host's cells (34). Viral proteins that have an experimentally identified and validated NoLS were used as an independent test of our human-trained predictor. As shown in Table 2, all NoLS-containing viral proteins considered were predicted to harbour at least one NoLS that overlaps with the experimentally validated NoLS.

Independent experimental validation of human proteins. The entire EVN dataset was encoded by considering both sequence and elements of structure and used to train an ANN which was then applied to the whole human proteome as defined by IPI version 3.40 (23). Supplementary File 5 shows the list of human proteins predicted to harbour a NoLS. The proteome-wide prediction of NoLSs may also be searched and downloaded from http://www.compbio.dundee.ac.uk/www-nod/.

Downloaded from http://nar.oxfordjournals.org/ at Universite de Sherbrooke on November 16, 2012

Table 2. Positions of experimentally validated and computationally predicted viral NoLSs

Protein name	Virus	Accession	Predicted NoLS position	Experimentally determined NoLS position	Reference for experimentally determined NoLS position
tat	HIV	NP_057853	43–68	48–61	(43)
rev	HIV	NP_057854	28–57	33–52	(44)
rex	HTLV-1	NP_057863	1–26	1–20	(45)
NS1A	Influenza A	P03495	208–237	216–237	(46)
US11	HSV1	NP_044674	86–105, 113–160	88–125	(47)
RL1	HSV1	P08353	1–22	1–16	(48)
ORF57	HVS	NP_040259	114–139	91–94, 119–128	(11,49)

Table 3. NoLSs chosen for experimental validation

Protein name	Accession	NoLS score	Subcellular localization of protein if known	Function/ process of protein if known	Reference for localization/ process annotations	NoLS cloned successfully	Experimentally validated as nucleolar-targeting
RBBP6	Q7Z6E9-4	0.981	N/A	N/A	N/A	Yes	Yes
RNF213	Q9HCF4-3	0.977	N/A	N/A	N/A	Yes	Yes
C1orf35	Q9BU76-1	0.976	N/A	N/A	N/A	Yes	Yes
DDX10	Q13206	0.970	N/A	RNA helicase	(50)	Yes	Yes
SF3B2	Q13435	0.966	Spliceosomal complex	RNA splicing	(51–53)	Yes	Yes
RBM34	P42696	0.966	Nucleolar (inferred from electronic annotation)	RNA binding (inferred from electronic annotation) (18)	(18)	No	N/A
CEBPZ	Q03701	0.959	Nucleus	Transcription	(54)	Yes	Yes
SMARCA2	P51531-2	0.958	Nucleoplasm	Regulation of transcription	(55)	Yes	Yes
AP3D1	O14617-4	0.957	Golgi apparatus	Intracellular protein transport	(56)	Yes	Yes
SRP72	O76094	0.950	Mainly cytoplasmic but nucleolar for complex assembly	Signal particle recognition binding	(57)	Yes	Yes
USP36	Q9P275-2	0.931	Nucleolar	Ubiquitin-dependent protein degradation (inferred from electronic annotation) (18)	(18,58)	N/A	Yes [independently validated (35)]

The predicted human NoLSs were ranked by score and ten of the highest scoring human NoLSs were chosen for experimental validation. Amongst the highest scoring NoLSs, care was taken to select diverse proteins including uncharacterized proteins (e.g. RNF213, C1orf35), mainly cytoplasmic proteins (AP3D1, SRP72), nucleoplasmic proteins (SMARCA2, CEBPZ) and a nucleolar protein for which no NoLS has been described (RBM34). These proteins selected for experimental validation are shown in Table 3 and the sequences of their NoLSs are shown in Table 4. Their respective high-scoring NoLSs were cloned downstream of GFP, expressed in U2OS cells and visualized by microscopy. GFP alone as well as a fusion protein of GFP cloned upstream of a region of protein RBM34 that is not predicted to be a NoLS (residues 324–345 of RBM34) were used as negative controls. As shown in Figure 4 and Supplementary File 6, all predicted NoLSs that were successfully cloned are capable of causing the accumulation of the GFP fusion protein in the nucleolus. The negative controls GFP and GFP-RBM34 (324–345) do not accumulate in the nucleolus. Interestingly, while all the predicted NoLS fusion proteins tested display a strong signal in the nucleolus, the extent of nucleoplasmic and cytoplasmic accumulations vary considerably for the different NoLSs. As the number of experimentally validated NoLSs increases in the future, it will become possible to investigate the differences between these signals and to determine whether they are NoLS-only or joint NoLS-NLS signals.

Table 4. Sequences of NoLSs chosen for experimental validation

Protein name	NoLS sequence chosen for experimental validation
RBBP6	SQDSKKKKKKEKKKHKKHKKHKKHKKH
RNF213	SWTVQESKKKKRKKKKKGNKSASSE
C1orf35	HRKSKKEKKKKKKRKHKKEKKKKDKEHRRP
DDX10	KKHSHRQNKKKQLRKQLKKPEWQVERE
SF3B2	GRSTVSVSKKEKNRKRRNRKKKKKPQRVRGVSSE
RBM34	KAVLLKTKKKGQKKSGRPKKQRKQK
CEBPZ	AKSIIKKKKHFKKKRIKTTQKTKKQRK
SMARCA2	QAQAAKEKKKRRRRKKKAEENAEGG
AP3D1	RRHRQKLEKDKRRKKRKEKEERTKGKKKSKK
SRP72	QPKEQGQGDLKKKKKKKKGKLPKNYDPK

In choosing the candidates for experimental validation, we also noticed that USP36 (described in Table 3), a high scoring candidate, has been recently validated by an

Downloaded from http://nar.oxfordjournals.org/ at Universite de Sherbrooke on November 16, 2012

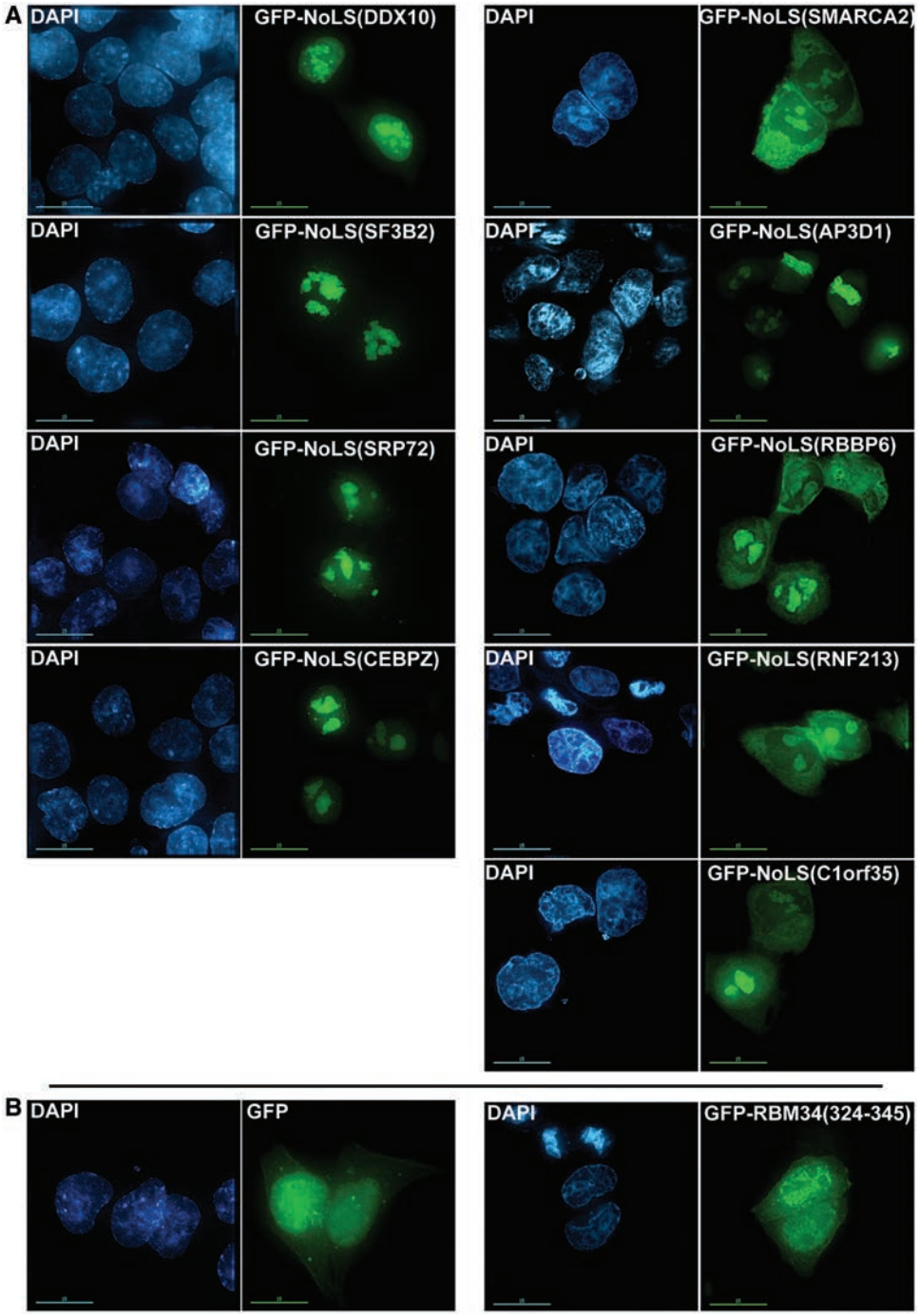

Figure 4. Experimental validation by microscopy. (**A**) Fusion constructs of NoLSs chosen for experimental validation and successfully cloned downstream of GFP (Table 3) were transfected into U2OS cells and the resulting proteins were visualized by microscopy [GFP-NoLS() labelled columns]. The DAPI columns show staining of the DNA in these cells. (B) GFP and GFP-RBM34(324–345) were used as negative controls. The bars represent 15 μm.

independent group. Endo and colleagues experimentally identified a functional NoLS between positions 1076 and 1091 of USP36 (35), while we predict an NoLS between residues 1073 and 1102.

Characteristics of NoLS-containing proteins

Analysis of whole-proteome predictions of NoLS reveals that a significantly larger proportion of proteins annotated as nucleolar are predicted to contain a NoLS than proteins annotated as localized in all other major cellular compartments (Figure 5). Of proteins annotated as nucleolar in Uniprot (18), 54% are predicted to harbour a NoLS. Thirty-nine percent of nuclear-annotated human proteins and 43% of nucleoplasmic or nuclear envelope human proteins are predicted to contain a NoLS. Since the nucleolus is contained within the nucleus, it is likely that many nucleolar proteins are still

Downloaded from http://nar.oxfordjournals.org/ at Universite de Sherbrooke on November 16, 2012

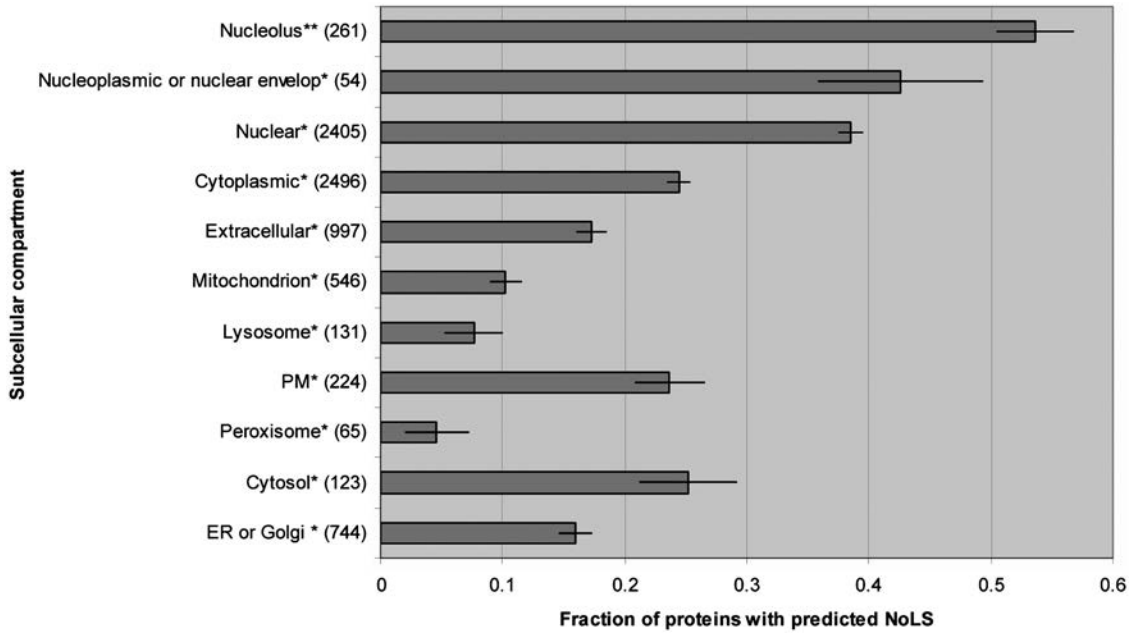

Figure 5. Characteristics of predicted NoLS-containing proteins. For all cellular compartments considered, the fraction of proteins predicted to harbour a NoLS is shown. Protein counts for each compartment are indicated in parenthesis beside the compartment name. The compartment groups labelled with an asterisk include proteins annotated as being in this and any other compartment except the nucleolus. The 261 proteins in the nucleolus group represent all proteins annotated as being nucleolar regardless of any other localization annotations they may have (indicated by double asterisks). The error bars were determined by bootstrap.

simply annotated as nuclear. As for the nucleoplasmic or nuclear envelope proteins predicted to have a NoLS, further experiments and a higher coverage of the localization annotations will be required to determine whether these proteins can also localize to the nucleolus or represent false-positive predictions. Amongst cytoplasmic proteins, between 25% (cytosolic proteins) and 5% (peroxisomal proteins) are predicted to contain NoLSs. While some of these proteins surely represent false-positive predictions, others are likely to represent true NoLS-containing proteins that might conditionally localize to the nucleolus. Numerous such examples have been reported (36–42).

In addition to the Uniprot localization annotations which are predominantly derived from microscopy experiments reported in the literature, we have also mapped our predictions of NoLSs onto the quantitative proteomic analysis of subcellular proteome localization described recently (10). In this study, the relative abundance of proteins in different cellular compartments was measured by harvesting nucleolar, nucleoplasmic and cytoplasmic cellular extracts each grown in the presence of amino acids labelled with different isotopes and then by pooling together the different fractions and analysing them by mass spectrometry. Table 5 shows the fraction of proteins that harbour at least one NoLS depending on their relative abundance ratios in the nucleolus. Similar to the Uniprot annotations, 48% of proteins that are both more nucleolar than nucleoplasmic and more nucleolar than cytoplasmic are predicted to harbour a NoLS. In contrast, ~25% of proteins that are more nucleoplasmic or cytoplasmic than nucleolar have a predicted NoLS and only 16% of proteins that are more nucleoplasmic and cytoplasmic than nucleolar harbour a predicted NoLS.

Significantly enriched Gene Ontology (GO) biological process annotations of all predicted NoLS-containing human proteins are shown in Table 6. The most prevalent terms associated with predicted NoLS-containing proteins involve transcription, processing of RNA and regulation of chromatin which agree well with the biological process annotations of many of the proteins that contain the EVN sequences.

DISCUSSION

NoLSs are emerging as a predominant mechanism in the targeting of proteins to the nucleolus. Through careful curation of the literature, we have identified 46 NoLSs, most of which are required for nucleolar targeting of the proteins that encode them and can target non-nucleolar reporter proteins to the nucleolus. As a group, these NoLSs contain a high proportion of basic amino acids making them similar to NLSs. Because of this similarity, NLSs and NoLSs are often perceived as analogous and interchangeably used to annotate proteins. In particular, short basic stretches in proteins are often assumed to be NLSs and even when experimental validation is performed, often no attention is given to the particular intra-nuclear localization of the protein even though this provides valuable clues about its function in the cell. Because of this, numerous NoLSs are annotated as NLSs.

Given the very different nature of their target compartments, the similarity between NLSs and NoLSs is

Downloaded from http://nar.oxfordjournals.org/ at Universite de Sherbrooke on November 16, 2012

Table 5. Comparison between NoLS predictions and protein localization ratios from ref. (10)

Localization abundance ratios	Total protein count	Protein count with predicted NoLSs	Fraction of proteins with NoLS (%)
Nucleolar/Cytoplasmic > 1 Nucleolar/Nucleoplasmic > 1	347	165	47.6
Nucleolar/Cytoplasmic ≤ 1 Nucleolar/Nucleoplasmic ≤ 1	1402	229	16.3
Nucleolar/Cytoplasmic ≤ 1 Nucleolar/Nucleoplasmic > 1	406	102	25.1
Nucleolar/Cytoplasmic > 1 Nucleolar/Nucleoplasmic ≤ 1	290	75	25.9

Table 6. Most significantly enriched GO annotations of predicted NoLS-containing proteins

Biological process GO term	Protein count[a]	Benjamini-adjusted *P*-value[b]	Fold enrichment[c]
GO:0006351~transcription, DNA-dependent	1008	5.42*E*−100	1.73
GO:0032774~RNA biosynthetic process	1008	1.23*E*−99	1.73
GO:0006355~regulation of transcription, DNA-dependent	988	1.03*E*−98	1.74
GO:0045449~regulation of transcription	1036	1.02*E*−98	1.71
GO:0051276~chromosome organization and biogenesis	221	2.10*E*−39	2.26
GO:0006323~DNA packaging	185	1.54*E*−35	2.34
GO:0006325~establishment and/or maintenance of chromatin architecture	181	1.37*E*−34	2.34
GO:0006259~DNA metabolic process	338	2.05*E*−30	1.76
GO:0016568~chromatin modification	120	1.23*E*−22	2.35
GO:0045934~negative regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	165	5.54*E*−20	1.97
GO:0016481~negative regulation of transcription	151	2.65*E*−18	1.97
GO:0031324~negative regulation of cellular metabolic process	174	7.51*E*−15	1.76
GO:0045892~negative regulation of transcription, DNA-dependent	111	5.30*E*−14	2.02
GO:0006333~chromatin assembly or disassembly	79	2.23*E*−13	2.28
GO:0008380~RNA splicing	113	1.27*E*−12	1.94
GO:0016071~mRNA metabolic process	140	1.51*E*−12	1.80

[a]Protein count of all predicted NoLS-containing proteins that are annotated with this GO term.
[b]The Benjamini-adjusted *P*-value was calculated by DAVID (25).
[c]Enrichment of this GO term in predicted NoLS-containing proteins compared to all human refseq proteins. Only GO terms with fold enrichment >1.7 are shown here.

somewhat surprising: NLSs specify translocation across the nuclear envelope, a double membrane surrounding the nucleus, whereas NoLSs ensure accumulation in the nucleolus, a membrane-less subcompartment within the nucleus. The similarity between NLSs and NoLSs has likely delayed the systematic characterization of NoLSs because of the extra difficulty of identifying clear and meaningful examples of both true NoLSs and true non-NoLSs. To overcome this problem, we have performed extensive curation of the literature making possible the accurate prediction of these motifs on a proteome-wide level. In future experiments, it will be important to consistently recognize and annotate NLSs and NoLSs as distinct, which will undoubtedly lead to improved predictions. A larger number of examples of true NLS-only signals, NoLS-only signals and joint NLS-NoLSs will help in better defining these signals and differentiating them. In addition to this, studies such as this one should help in the construction of precisely targeted fusion proteins, ensuring that proteins are not highly enriched in the nucleolus when the aim is to locate them in the nucleoplasm.

A small number of proteins have been proposed to act as transporters to the nucleolus [e.g. B23/NPM1 which shuttles between the cytoplasm and nucleolus and binds several NoLS-containing proteins (28)]. Alternatively, NoLSs might instead bind to nucleolar RNA thus causing the targeting of the proteins that contain them to the nucleolus. Further investigations will be required to clarify whether protein transporters are widely used for the nucleolar targeting of NoLS-containing proteins or whether other mechanisms are predominantly employed for this purpose. The NoLS predictions should serve as a good starting point to experimentally address these questions.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We would like to thank Drs Tom Walsh and Peter Troshin for technical expertise.

FUNDING

M.S.S. is a recipient of a post-doctoral fellowship from the Caledonian Research Foundation. A.I.L. is a Wellcome

Downloaded from http://nar.oxfordjournals.org/ at Universite de Sherbrooke on November 16, 2012

Trust Principal Research Fellow. A.I.L. and F.M.B. are funded in part by the European Commission's FP7 (GA HEALTH-F4-2008-201648/PROSPECTS) (www.prospects-fp7.eu/) and by a Wellcome Trust programme grant (073980/Z/03/Z). G.J.B. acknowledges funding from the Wellcome Trust (WT083481). Funding for open access charge: Wellcome Trust grant WT083481.

Conflict of interest statement. None declared.

REFERENCES

1. Scheer,U. and Hock,R. (1999) Structure and function of the nucleolus. Curr. Opin. *Cell Biol.*, **11**, 385–390.
2. Boisvert,F.M., van Koningsbruggen,S., Navascues,J. and Lamond,A.I. (2007) The multifunctional nucleolus. *Nat. Rev. Mol. Cell Biol.*, **8**, 574–585.
3. Olson,M.O., Dundr,M. and Szebeni,A. (2000) The nucleolus: an old factory with unexpected capabilities. *Trends Cell Biol.*, **10**, 189–196.
4. Olson,M.O., Hingorani,K. and Szebeni,A. (2002) Conventional and nonconventional roles of the nucleolus. *Int. Rev. Cytol.*, **219**, 199–266.
5. Pederson,T. (1998) The plurifunctional nucleolus. *Nucleic Acids Res.*, **26**, 3871–3876.
6. Pederson,T. and Tsai,R.Y. (2009) In search of nonribosomal nucleolar protein function and regulation. *J. Cell Biol.*, **184**, 771–776.
7. Pederson,T. (1998) Growth factors in the nucleolus? *J. Cell Biol.*, **143**, 279–281.
8. Ahmad,Y., Boisvert,F.M., Gregor,P., Cobley,A. and Lamond,A.I. (2009) NOPdb: Nucleolar Proteome Database–2008 update. *Nucleic Acids Res.*, **37**, D181–184.
9. Andersen,J.S., Lam,Y.W., Leung,A.K., Ong,S.E., Lyon,C.E., Lamond,A.I. and Mann,M. (2005) Nucleolar proteome dynamics. *Nature*, **433**, 77–83.
10. Boisvert,F.M., Lam,Y.W., Lamont,D. and Lamont,A.I. (2010) A quantitative proteomic analysis of subcellular proteome localization and changes induced by DNA damage. *Mol. Cell Proteomics*, **9**, 457–470.
11. Emmott,E. and Hiscox,J.A. (2009) Nucleolar targeting: the hub of the matter. *EMBO Rep.*, **10**, 231–238.
12. Boulikas,T. (1993) Nuclear localization signals (NLS). *Crit. Rev. Eukaryot. Gene Expr.*, **3**, 193–227.
13. von Heijne,G. (1990) The signal peptide. *J. Membr. Biol.*, **115**, 195–201.
14. Gavel,Y., Nilsson,L. and von Heijne,G. (1988) Mitochondrial targeting sequences. Why 'non-amphiphilic' peptides may still be amphiphilic. *FEBS Lett.*, **235**, 173–177.
15. Gould,S.J., Keller,G.A., Hosken,N., Wilkinson,J. and Subramani,S. (1989) A conserved tripeptide sorts proteins to peroxisomes. *J. Cell Biol.*, **108**, 1657–1664.
16. Carmo-Fonseca,M., Mendes-Soares,L. and Campos,I. (2000) To be or not to be in the nucleolus. *Nat. Cell Biol.*, **2**, E107–E112.
17. Nair,R., Carter,P. and Rost,B. (2003) NLSdb: database of nuclear localization signals. *Nucleic Acids Res.*, **31**, 397–399.
18. The Universal Protein Resource. (2010) (UniProt) in 2010. *Nucleic Acids Res.*, **38**, D142–D148.
19. Hobohm,U., Scharf,M., Schneider,R. and Sander,C. (1992) Selection of representative protein data sets. *Protein Sci.*, **1**, 409–417.
20. Nielsen,H., Engelbrecht,J., von Heijne,G. and Brunak,S. (1996) Defining a similarity threshold for a functional protein sequence pattern: the signal peptide cleavage site. *Proteins*, **24**, 165–177.
21. Pearson,W.R. and Lipman,D.J. (1988) Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.
22. Sigrist,C.J., Cerutti,L., de Castro,E., Langendijk-Genevaux,P.S., Bulliard,V., Bairoch,A. and Hulo,N. (2010) PROSITE, a protein domain database for functional characterization and annotation. *Nucleic Acids Res.*, **38**, D161–D166.
23. Kersey,P.J., Duarte,J., Williams,A., Karavidopoulou,Y., Birney,E. and Apweiler,R. (2004) The International Protein Index: an integrated database for proteomics experiments. *Proteomics*, **4**, 1985–1988.
24. Cole,C., Barber,J.D. and Barton,G.J. (2008) The Jpred 3 secondary structure prediction server. *Nucleic Acids Res.*, **36**, W197–W201.
25. Huang da,W., Sherman,B.T. and Lempicki,R.A. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.*, **4**, 44–57.
26. Cokol,M., Nair,R. and Rost,B. (2000) Finding nuclear localization signals. *EMBO Rep.*, **1**, 411–415.
27. Rubtsov,Y.P., Zolotukhin,A.S., Vorobjev,I.A., Chichkova,N.V., Pavlov,N.A., Karger,E.M., Evstafieva,A.G., Felber,B.K. and Vartapetian,A.B. (1997) Mutational analysis of human prothymosin alpha reveals a bipartite nuclear localization signal. *FEBS Lett.*, **413**, 135–141.
28. Valdez,B.C., Perlaky,L., Henning,D., Saijo,Y., Chan,P.K. and Busch,H. (1994) Identification of the nuclear and nucleolar localization signals of the protein p120. Interaction with translocation protein B23. *J. Biol. Chem.*, **269**, 23776–23783.
29. Liu,J., Du,X. and Ke,Y. (2006) Mapping nucleolar localization sequences of 1A6/DRIM. *FEBS Lett.*, **580**, 1405–1410.
30. Huang,H.S., Pozarowski,P., Gao,Y., Darzynkiewicz,Z. and Lee,E.Y. (2005) Protein phosphatase-1 inhibitor-3 is co-localized to the nucleoli and centrosomes with PP1gamma1 and PP1alpha, respectively. *Arch. Biochem. Biophys.*, **443**, 33–44.
31. Baldi,P. and Brunak,S. (2001) *Bioinformatics: The Machine Learning Approach*, 2nd edn. MIT Press, Cambridge, MA.
32. Nielsen,H., Engelbrecht,J., Brunak,S. and von Heijne,G. (1997) Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng.*, **10**, 1–6.
33. Prieve,M.G., Guttridge,K.L., Munguia,J. and Waterman,M.L. (1998) Differential importin-alpha recognition and nuclear transport by nuclear localization signals within the high-mobility-group DNA binding domains of lymphoid enhancer factor 1 and T-cell factor 1. *Mol. Cell Biol.*, **18**, 4819–4832.
34. Hiscox,J.A. (2007) RNA viruses: hijacking the dynamic nucleolus. *Nat. Rev. Microbiol.*, **5**, 119–127.
35. Endo,A., Kitamura,N. and Komada,M. (2009) Nucleophosmin/B23 regulates ubiquitin dynamics in nucleoli by recruiting deubiquitylating enzyme USP36. *J. Biol. Chem.*, **284**, 27918–27923.
36. Dang,C.V. and Lee,W.M. (1989) Nuclear and nucleolar targeting sequences of c-erb-A, c-myb, N-myc, p53, HSP70, and HIV tat proteins. *J. Biol. Chem.*, **264**, 18019–18023.
37. Henderson,J.E., Amizuka,N., Warshawsky,H., Biasotto,D., Lanske,B.M., Goltzman,D. and Karaplis,A.C. (1995) Nucleolar localization of parathyroid hormone-related peptide enhances survival of chondrocytes under conditions that promote apoptotic cell death. *Mol. Cell Biol.*, **15**, 4064–4075.
38. Stegh,A.H., Schickling,O., Ehret,A., Scaffidi,C., Peterhansel,C., Hofmann,T.G., Grummt,I., Krammer,P.H. and Peter,M.E. (1998) DEDD, a novel death effector domain-containing protein, targeted to the nucleolus. *Embo J.*, **17**, 5974–5986.
39. Caron,E., Cote,C., Parisien,M., Major,F. and Perreault,C. (2006) Identification of two distinct intracellular localization signals in STT3-B. *Arch. Biochem. Biophys.*, **445**, 108–114.
40. Stark,L.A. and Dunlop,M.G. (2005) Nucleolar sequestration of RelA (p65) regulates NF-kappaB-driven transcription and apoptosis. *Mol. Cell Biol.*, **25**, 5985–6004.
41. Antoine,M., Reimers,K., Dickson,C. and Kiefer,P. (1997) Fibroblast growth factor 3, a protein with dual subcellular localization, is targeted to the nucleus and nucleolus by the concerted action of two nuclear localization signals and a nucleolar retention signal. *J. Biol. Chem.*, **272**, 29475–29481.
42. Goyal,P., Pandey,D. and Siess,W. (2006) Phosphorylation-dependent regulation of unique nuclear and nucleolar localization signals of LIM kinase 2 in endothelial cells. *J. Biol. Chem.*, **281**, 25223–25230.
43. Siomi,H., Shida,H., Maki,M. and Hatanaka,M. (1990) Effects of a highly basic region of human immunodeficiency virus Tat protein on nucleolar localization. *J. Virol.*, **64**, 1803–1807.

Downloaded from http://nar.oxfordjournals.org/ at Universite de Sherbrooke on November 16, 2012

44. Bohnlein,E., Berger,J. and Hauber,J. (1991) Functional mapping of the human immunodeficiency virus type 1 Rev RNA binding domain: new insights into the domain structure of Rev and Rex. *J. Virol.*, **65**, 7051–7055.
45. Nosaka,T., Siomi,H., Adachi,Y., Ishibashi,M., Kubota,S., Maki,M. and Hatanaka,M. (1989) Nucleolar targeting signal of human T-cell leukemia virus type I rex-encoded protein is essential for cytoplasmic accumulation of unspliced viral mRNA. *Proc. Natl Acad. Sci. USA*, **86**, 9798–9802.
46. Melen,K., Kinnunen,L., Fagerlund,R., Ikonen,N., Twu,K.Y., Krug,R.M. and Julkunen,I. (2007) Nuclear and nucleolar targeting of influenza A virus NS1 protein: striking differences between different virus subtypes. *J. Virol.*, **81**, 5995–6006.
47. Catez,F., Erard,M., Schaerer-Uthurralt,N., Kindbeiter,K., Madjar,J.J. and Diaz,J.J. (2002) Unique motif for nucleolar retention and nuclear export regulated by phosphorylation. *Mol. Cell Biol.*, **22**, 1126–1139.
48. Cheng,G., Brett,M.E. and He,B. (2002) Signals that dictate nuclear, nucleolar, and cytoplasmic shuttling of the gamma(1)34.5 protein of herpes simplex virus type 1. *J. Virol.*, **76**, 9434–9445.
49. Boyne,J.R. and Whitehouse,A. (2006) Nucleolar trafficking is essential for nuclear export of intronless herpesvirus mRNA. *Proc. Natl Acad. Sci. USA*, **103**, 15190–15195.
50. Savitsky,K., Ziv,Y., Bar-Shira,A., Gilad,S., Tagle,D.A., Smith,S., Uziel,T., Sfez,S., Nahmias,J., Sartiel,A. *et al.* (1996) A human gene (DDX10) encoding a putative DEAD-box RNA helicase at 11q22-q23. *Genomics*, **33**, 199–206.
51. Gozani,O., Feld,R. and Reed,R. (1996) Evidence that sequence-independent binding of highly conserved U2 snRNP proteins upstream of the branch site is required for assembly of spliceosomal complex A. *Genes Dev.*, **10**, 233–243.
52. Neubauer,G., King,A., Rappsilber,J., Calvio,C., Watson,M., Ajuh,P., Sleeman,J., Lamond,A. and Mann,M. (1998) Mass spectrometry and EST-database searching allows characterization of the multi-protein spliceosome complex. *Nat. Genet.*, **20**, 46–50.
53. Zhou,Z., Licklider,L.J., Gygi,S.P. and Reed,R. (2002) Comprehensive proteomic analysis of the human spliceosome. *Nature*, **419**, 182–185.
54. Lum,L.S., Sultzman,L.A., Kaufman,R.J., Linzer,D.I. and Wu,B.J. (1990) A cloned human CCAAT-box-binding factor stimulates transcription from the human hsp70 promoter. *Mol. Cell Biol.*, **10**, 6709–6717.
55. Muchardt,C., Reyes,J.C., Bourachot,B., Leguoy,E. and Yaniv,M. (1996) The hbrm and BRG-1 proteins, components of the human SNF/SWI complex, are phosphorylated and excluded from the condensed chromosomes during mitosis. *EMBO J.*, **15**, 3394–3402.
56. Simpson,F., Peden,A.A., Christopoulou,L. and Robinson,M.S. (1997) Characterization of the adaptor-related protein complex, AP-3. *J. Cell Biol.*, **137**, 835–845.
57. Politz,J.C., Yarovoi,S., Kilroy,S.M., Gowda,K., Zwieb,C. and Pederson,T. (2000) Signal recognition particle components in the nucleolus. *Proc. Natl Acad. Sci. USA*, **97**, 55–60.
58. Barbe,L., Lundberg,E., Oksvold,P., Stenius,A., Lewin,E., Bjorling,E., Asplund,A., Ponten,F., Brismar,H., Uhlen,M. *et al.* (2008) Toward a confocal subcellular atlas of the human proteome. *Mol. Cell Proteomics*, **7**, 499–508.

Downloaded from http://nar.oxfordjournals.org/ at Universite de Sherbrooke on November 16, 2012