Abstract: Although proteins are translated on cytoplasmic ribosomes, many of these proteins play essential roles in the nucleus, mediating key cellular processes including but not limited to DNA replication and repair as well as transcription and RNA processing. Thus, understanding how these critical nuclear proteins are accurately targeted to the nucleus is of paramount importance in biology. Interaction and structural studies in the recent years have jointly revealed some general rules on the specificity determinants of the recognition of nuclear targeting signals by their specific receptors, at least for two nuclear import pathways: (i) the classical pathway, which involves the classical nuclear localization sequences (cNLSs) and the receptors importin-\(\alpha\)/karyopherin-\(\alpha\) and importin-\(\beta\)/karyopherin-\(\beta1\); and (ii) the karyopherin-\(\beta2\) pathway, which employs the proline-tyrosine (PY)-NLSs and the receptor transportin-1/karyopherin-\(\beta2\). The understanding of specificity rules allows the prediction of protein nuclear localization. We review the current understanding of the molecular determinants of the specificity of nuclear import, focusing on the importin-\(\alpha\)-cargo recognition, as well as the currently available databases and predictive tools relevant to nuclear localization.
Molecular basis for specificity of nuclear import and prediction of nuclear localization

Mary Marfori a, Andrew Mynott b, Jonathan Ellis a, Ahmed Mehdi c, Neil F. W. Saunders a,1, Paul M. Curmi b, Jade K. Forwood d, Mikael Boden c, Bostjan Kobe a,c,*

a School of Chemistry and Molecular Biosciences and Centre for Infectious Disease Research, University of Queensland, Brisbane, Queensland 4072, Australia.
b School of Physics, University of New South Wales, Sydney, NSW 2052, Australia.
c Institute for Molecular Bioscience, University of Queensland, Brisbane, Queensland 4072, Australia.
d School of Biomedical Sciences, Charles Sturt University, Wagga Wagga, New South Wales 2650, Australia.

1 Current address: CSIRO Mathematics, Informatics and Statistics, Locked Bag 17, North Ryde, NSW 1670, Australia

*Correspondence: Bostjan Kobe, phone +61733652132, fax +61733654699, e-mail b.kobe@uq.edu.au.
ABSTRACT

Many proteins have functions in the nucleus, and therefore the understanding of the processes of nucleo-cytoplasmic transport and the determinants of nuclear targeting signals is extremely important. Interaction and structural studies in the recent years have jointly revealed some general rules on the specificity determinants of the recognition of nuclear targeting signals by their specific receptors, at least for two nuclear import pathways: (i) the classical pathway, which involves the classical nuclear localization sequence (cNLS) and the receptors importin-α/karyopherin-α and importin-β/karyopherin-β1; and (ii) the karyopherin-β2 pathway, which employs the proline-tyrosine (PY)-NLSs and the receptor karyopherin-β2/transportin-1. The understanding of specificity rules will allow the prediction of protein localization. We review the current understanding of the molecular determinants of the specificity of nuclear import, as well as bioinformatic tools currently available to predict nuclear localization.

Keywords: importin-α/karyopherin-α; importin-β/karyopherin-β1; nuclear localization; nuclear localization sequence (NLS); nucleocytoplasmic transport; prediction of nuclear localization; transportin-1/karyopherin-β2
1. Overview of nuclear transport pathways and the determinants of nuclear localization

The nucleus is the defining feature of a eukaryotic cell. It allows the segregation of the genetic material and the transcriptional machinery (in the nucleus) from the translational and metabolic machinery (in the cytoplasm). This segregation facilitates the regulation of diverse cellular processes including gene expression, signal transduction and the cell cycle. One of the key features of such regulation is the selective bi-directional transport between the cytoplasm and the nucleus. Such transport occurs through large membrane structures termed the nuclear pore complexes (NPCs), comprised of ~30 proteins collectively termed nucleoporins (Nups) [1, 2]. The mechanism of the translocation through the pore is not well understood, but is facilitated by a number of nuclear transport factors or carrier proteins termed karyopherins. Many of these factors are members of a superfamily of proteins containing HEAT repeats and collectively called β-karyopherins; there are over 20 β-karyopherins present in human cells, and 14 in budding yeast [3-5] (see also article by Chook et al in this issue). While macromolecules smaller than ~40 kDa can passively diffuse through the pore, most if not all proteins with functions in the nucleus use active carrier-mediated transport.

The transport processes are generally regulated by the small GTPase protein Ran. Ran cycles between GDP and GTP-bound states. The nature of this nucleotide-bound state is in turn modulated by regulatory proteins, the most important being Ran guanine nucleotide exchange factor (RanGEF; regulator of chromosome condensation 1, RCC1) and the Ran GTPase-activating protein (RanGAP). Because RanGEF is compartmentalized to the nucleus, and RanGAP to the cytoplasm, there is an asymmetric distribution of the different nucleotide-bound forms of Ran. This RanGDP-RanGTP gradient from the cytoplasm to the nucleus results in a directionality of nucleo-cytoplasmic pathways. In general, import receptors bind cargo in the cytoplasm, and release it in the nucleus upon RanGTP binding. Conversely, export receptors bind cargo on the nucleus in complex with RanGTP, and the dissociation is triggered upon GTP hydrolysis.

The main determinant of nuclear localization is the nuclear targeting signal. The best-characterized nuclear targeting signal is the classical nuclear localization sequence (cNLS), which is recognized by the protein importin-α (karyopherin-α; Impα). Impα is an adaptor protein, because it requires in turn binding to importin-β (karyopherin-β 1; Impβ1), and it is the trimeric Impα•Impβ1•cNLS-containing cargo complex that can enter the nucleus. While this pathway is termed the classical nuclear import pathway, several alternative import pathways have been characterized. The localization of the protein is further influenced by the presence of a nuclear export signal (NES [6, 7]; many proteins can shuttle in and out of the nucleus), the size of the protein and many other factors.

Understanding the determinants of specificity of nuclear targeting signals requires the integration of structural and interaction data on receptor•signal binding, for a number of natural and designed variants of a targeting signal. In this review, we focus on the determinants of cNLSs as the best characterized example. Such information can be used
to predict nuclear localization of uncharacterized proteins. For example, similar information has been used successfully to develop a bioinformatic tool to predict sequences phosphorylated by eukaryotic kinases (Predikin; [8, 9]), which has shown the best performance among prediction tools at the recent DREAM4 Peptide Recognition Domain Specificity Prediction [10]. In the case of nuclear targeting signals, such information has not yet been extensively used for prediction of nuclear localization, although a number of alternative approaches have been used to enable such predictions. We review the predictions methods and nuclear localization data resources available currently.

In this paper, we first review different nuclear import pathways and the associated determinants of specificity, focusing on the best-characterized classical nuclear import pathway. We then review bioinformatic resources of nuclear localization data and for prediction of nuclear localization.

2. Nuclear import pathways and the associated determinants of specificity

2.1. The classical nuclear import pathway

The first nuclear targeting signal has been identified through mutational studies of the Simian virus 40 (SV40) large T-antigen (TAg) [11, 12]. A related signal has subsequently been identified in *Xenopus* nucleoplasmin [13], and both these two sequences are now classified as cNLSs and were found to require the proteins Impα and Impβ1 for nuclear transport [14-16] (for a recent review, see [17]).

2.1.1. Classical NLSs are recognized by the adaptor protein importin-α

Impα is comprised of two functionally and structurally distinct domains: the N-terminal Impβ1 binding (αIBB) domain, and a C-terminal domain that consists of ten tandem armadillo (Arm) repeats [18-22] (see also the article by Cingolani in this issue). The Arm repeat consists of three α-helices (termed H1, H2 and H3), encoded by a ~40 amino acid motif, first identified in the *Drosophila* armadillo protein [23]. The consecutive stacking of Arm repeats generates a superhelical structure, forming an elongated banana-shaped molecule (Fig. 1). The H3 helices define the inner concave surface of the protein. Structures of yeast, mouse and human Impα have been elucidated by x-ray crystallography, all displaying analogous superhelical architectures [24-27].

2.1.2. Structural basis of cNLS recognition: the major and minor cNLS-binding sites

The cNLSs contain one or two stretches of basic amino acids (therefore they are grouped into two classes, monopartite and bipartite NLSs, respectively) [28]. Structural studies (Table 1) reveal that the NLS binding site is located along a groove on Impα surface that shows a high degree of conservation (Fig. 1C). This groove contains an array of conserved tryptophan and asparagine residues at the third and fourth turns of the H3 helices. This array is disrupted in Arm repeats 5 and 6, segregating the binding groove into the “major” (comprised of residues from Arm repeats 2-4) and the “minor” site (comprised of residues from Arm repeats 6-8). The disruption in the tryptophan-asparagine array is conserved between yeast, mouse and human Impα proteins [24, 26, 27]. Bipartite cNLSs bind in an extended conformation to both major and minor binding
sites, with the N-terminal basic cluster in the minor site, and the C-terminal basic cluster in the major site. The main chain of the peptide therefore runs antiparallel to the direction of the Impα superhelix. In this conformation, the cNLS backbone can form favourable contacts with the appropriately spaced conserved asparagine residues. Monopartite cNLSs bind in an analogous manner, and can bind to either of the two sites. The major site is considered to be the high-affinity binding site for typical monopartite cNLSs.

Distinct hydrophobic grooves are formed along the cNLS binding site as a result of the indole-stacking array of the conserved tryptophan residues, accommodating the extended aliphatic fractions of the basic side-chains in the cNLS. The positively charged portions can subsequently form salt bridge and hydrogen bonding interactions with strategically positioned negatively charged residues that line the binding pockets, and also interact with the helix dipoles of the H3 helices.

The major and minor cNLS binding sites have also been shown to interact with a segment of the αIBB domain [29], and to the nucleoporin Nup50 (Npap60, Nup2p in yeast) [30, 31], respectively. The former interaction is thought to have an autoinhibitory function [29, 32, 33], while the latter has been proposed to accelerate cargo release in the nucleus [30, 34].

Structural studies of a number of peptides and proteins in complex with yeast, mouse and human Impαs reveal a highly conserved mechanism of recognition [25-27, 35-41] (Table 1). Integration of structural studies and interaction studies of variants of NLSs (e.g. [42]) jointly helps us understand the specificity of cNLS binding.

The critical residues in a bipartite cNLS have been termed P1’-P2’ and P2-P5 for the N- and C-terminal basic clusters, respectively [37] (Fig. 1D). Table 1 shows a structural alignment of all cNLS-like sequences for which the binding to Impα has been structurally characterized. The most critical binding pocket appears to be P2, predominantly defined by Thr155 and Asp192 (Thr166, Asp203 in yeast) in Arm repeats 2 and 3 of the mouse adaptor protein. The dimensions of the pocket appear best suited for binding a lysine residue. In the monopartite SV40 TAg NLS, mutagenesis of the P2 lysine to non-basic residues abolished nuclear localisation of the protein [11]. The K128A mutation resulted in a ~300 fold decrease in affinity towards Impα compared to the wildtype cNLS [42]. The larger arginine residue at this position appears to be energetically less favoured; the K128R substitution in the SV40 TAg NLS produces a ~ 2.7 kcal/mol decrease in binding free energy [42, 43].

Appropriately, this P2 lysine is strictly conserved among all structurally identified bipartite cNLSs (Table). Structurally, the necessity of the P2 lysine is rationalized through the specific and extensive hydrogen bonding interactions with the adaptor protein at this position. The terminal nitrogen atom of the lysine side chain coordinates with the main chain carbonyl group of Gly150, the hydroxyl side chain of Thr155, and to the negatively charged side chain of Asp192. Structures suggest that a longer arginine side chain at this position would be unable to maintain the favourable side chain hydrogen bonding arrangement and may also force the cNLS main chain to adopt an unfavourable position.
Positions P3 and P5 of the major site are also well defined, displaying a preference for long basic side chains such as arginine or lysine. The relative energy contributions to cNLS binding at these positions are approximately equal, though only 2/3 of the contribution of lysine at P2 [42]. Mutational and peptide library studies [41] for the P3 position suggest a preference for an arginine residue over that of a lysine. A structural interpretation of this binding register reveals that the longer arginine side chain would be able to participate in more favourable electrostatic interactions with Glu266/Asp270 (Glu272 and Asp276 in yeast) of the binding groove. However, there is a lack of hydrogen bonding interactions at these positions contributing to a less discriminate binding specificity (Table 1; αIBB domain, PLSCR1 and AR).

The P4 binding pocket is notably more tolerant to accepting a diverse range of residues than P2, P3 and P5 pockets. Consistently, the energy contribution from this pocket is only ~1/4 of the contribution of the P2 position [42]. The proportions of the groove suggest that smaller hydrophobic residues could easily be accommodated in the P4 position. Several structures show a smaller hydrophobic and hydrophilic residue in this position (Table 1 PLSCR1, PB2, RB, N1N2, c-Myc), further suggesting that a basic residue is not critical in this pocket. Nonetheless, an arginine would make the most favourable interactions with the H3 helix dipole of ARM repeats 1 and 2, and this weak electrostatic interaction provides some explanation for the slight preference for a basic residue at this position.

An inspection of the P1 and P6 binding pockets suggests that there is little preference at these positions, consistent with mutagenesis studies [11]. There are few side chain interactions at these locations that could provide specificity. Accordingly, no distinct trend is seen for either pocket, with an assortment of side chains of different size and properties shown to bind at these positions (Table 1).

The binding pockets for the individual side chains are less well defined in the minor site compared to the major site. Energetically, this site contributes to only a modest increase in free binding (~3 kcal/mol), approximately equal to that of the binding energy for a single residue at P3 or P5 in the major site [42]. In all bipartite cNLS•Impα structures characterised to date, a lysine-arginine motif is present in positions P1′ and P2′ of the minor site, respectively. Mutations of this motif to other (non- Lys/Arg) amino-acids have resulted in the cytoplasmic accumulation of various bipartite cNLS-containing proteins [44, 45]. The P1′ pocket is defined by residues Thr328 and Asp361 (Thr322 and Asn367 in yeast). A lysine is preferred in this binding cavity, as an arginine side chain at this position is too long to make favourable interactions with the Impα side chains. The P2′ binding pocket is defined within the tryptophan stacking array of Arm repeats 7 and 8, with conserved residues Glu396 and Ser360 (Glu402 and Ser366 in yeast) located at the base of the groove. A lysine can be accommodated at this position, as seen in the monopartite SV40 TAg structure. However, an arginine side chain is able to make more favourable contacts than a lysine at this position.

2.1.3. Atypical cNLSs

One of the atypical cNLSs with a known structure in complex with Impα is the hydrophobic cNLS from phospholipid scramblase 1 (PLSCR1), which nevertheless binds at the major cNLS binding site analogous to other monopartite cNLSs [38] (Table 1). The
binding mode may shed light on some other atypical Impα-binding cNLSs such as the one from the Borna disease virus p10 protein (6RLTLLELVRRLNGN19) [46].

Recent studies have identified some atypical NLSs that primarily bind to the minor site of Impα. The structure of TPX2 (target protein for Xenopus kinesin-like protein 2) cNLS bound to mouse Impα shows the standard KR motif in positions P1' and P2', respectively [40]. However, the peptide achieves more extensive interactions at the minor site of the importin protein through exploiting positions P3' and P4'. The lysine in P3' position cavity participates in hydrogen bonding interactions with Impα Gly281, Asn283 and Thr322. Notably, the histidine at P4' makes a novel interaction with Impα Trp357 and a salt bridge with Glu354. The cNLSs from STAT1 (signal transducers and activators of transcription 1) protein and the influenza A virus NP (nucleoprotein) similarly bind preferentially to the minor binding site of Impα, as inferred from mutagenesis of the Impα binding pockets [47, 48]. The mechanism of the STAT1•Impα interaction is particularly unusual, requiring the homodimerization of the STAT1 protein to constitute a functional cNLS, with the basic cNLS residues contributed from both subunits [48].

A study using random peptide libraries identified two groups of monopartite cNLSs that were reported to bind specifically to the minor site of Impα [49]. Although the structural basis of binding of these peptides to Impα has not been determined, one of the groups closely resembles the sequence of TPX2, with the exception of a hydrophobic residue at P4'.

2.1.4. Linker and flanking regions of cNLSs

The linker region between the two basic clusters in bipartite cNLSs is usually 10-12 residues long and often makes no specific contacts with the adaptor protein, consistent with the ability of this sequence to tolerate mutations without abolishing nuclear localisation. Specific interactions found occasionally in the linker region are not conserved and appear to be specific to a particular cNLS. The linker sequence needs to span about ten residues, so that the basic clusters can bind favourably at both major and minor sites in an extended conformation.

A recent study of the effect of linker sequences on Impα binding and nuclear import showed that significantly longer linker sequences can be functional [50]. Relevant to this observation, the structure of TPX2 cNLS•Impα complex revealed residues 327KMIK330 bound the major site of Impα, in addition to the sequence bound to the minor site [40]. While this major site interaction was shown not to be critical for TPX2 to bind to the adaptor protein, this basic cluster is 40 amino acids downstream from the minor site-binding cluster of TPX2. The effect of the linker region was also investigated through activity-based profiling via systematic mutational analysis of a bipartite cNLS [51]. The study found that acidic residues in the linker regions, as well as further basic residues flanking the basic clusters N- and C-terminally, respectively, can contribute significantly to the interaction [51]. The structural basis of these effects is not currently known.

The sequences flanking the basic clusters have also been shown in other studies to contribute to the cNLS•Impα interaction, for example in the case of SV40 TAg [52, 53]. Furthermore, phosphorylation of this region further enhances the affinity and provides a means to regulate nuclear localization. Interestingly, the crystal structure of the
corresponding phosphorylated cNLS showed no interaction of the phosphate with Impα, therefore the basis of the enhanced affinity remains unclear [54]. However, the flanking sequence was found to form a number of favourable interactions with the Impα surface distinct to that interacting with the linker region in bipartite cNLSs. Together, the available studies suggest that the bipartite cNLS linker and flanking regions can differentially exploit the Impα surface in different cNLSs and enhance the cNLS•adaptor protein interaction.

2.1.5. Interactions of cNLSs with importin-α in the context of the native proteins

Classical NLSs sequences are usually found at N- and C-termini of proteins, between domains or in flexible loop regions. This is consistent with the mode of interaction revealed by structural studies, suggesting that cNLSs adopt an ordered structure only upon binding Impα, in line with the expected mechanism of action of the majority of linear motifs [55]. Most of the available structural studies have therefore used peptides corresponding to cNLSs, outside of the context of the native protein. However, two recent studies reported successful crystallization of Impα bound to a cNLSs in the context of the native protein; these involve the structure of human Impα5 bound to a 82-residue C-terminal domain fragment of influenza virus polymerase PB2 (residues 678-759) [27], and the structure of human Impα1 bound to cap-binding protein 80 (CBP80) [26]. No interactions outside the bipartite CBP80 cNLS were identified in the latter study. However, in the case of PB2, a lysine residue 20 amino-acids N-terminal to the bipartite cNLS was found to hydrogen-bond to residues Gly284, Asn286 and Thr325 on the surface of Impα5 (residues Gly281, Asn283 and Thr322 in mouse Impα) [27]. These residues form part of the P3′ pocket, an important binding site for TPX2 [40].

2.1.6. Determinants of cNLS specificity

While many differences exist between specific cNLSs, common features clearly emerge that allow us to develop some general rules that underpin cNLS binding. The deduction of these rules hinges on the integration of structural and interaction studies. Peptide library studies and systematic mutagenesis studies provide a complementary approach to identifying these general rules, and several such studies have been carried out recently in this system. An oriented peptide library approach probed the specificity of binding to the major cNLS binding site [41]; random peptide library screening using mRNA display selected a range of monopartite and bipartite cNLS-like sequences [49]; and an additivity-based design using the activity-based profile derived via extensive systematic mutagenesis of a bipartite NLS yielded high-affinity Impα-binding peptides [51]. The various analyses generally agree; examples of proposed optimal consensus sequences include K(R/K)X(R/K) [56], K(K/R)X(K/R) [42], KR(R/X)K [37], KRRR [51] and KR(K/R)R or K(K/R)RK [49] for a monopartite cNLS, and (K/R)(K/R)X_{10-12}K(R/R)_{3/5} [44], KR_{10-12}KRRK [54] and KR_{10-12}K(K/R)(K/R) and KR_{10-12}K(K/R)X(K/R) [49] for a bipartite cNLS. In a bipartite cNLS, the constraints for the major site binding by the C-terminal cluster are relaxed compared to a monopartite cNLS, because of the additional contributions to the affinity by the minor site binding. It now emerges that longer linker
lengths can also be found in functional cNLSs. Further studies will be required to refine our understanding of the rules governing the binding to the minor site. It is becoming clear that a simple consensus sequence cannot adequately capture all the properties of a cNLS, and a more complex description using a position-specific scoring matrix (PSSM) for example is required to better capture the properties of a functional cNLS [57] (see Section 3 below). Finally, it is clear that specific cNLSs often use interactions that do not conform to the general rules, and this complicates a general definition of a cNLS.

In animals, duplications and expansions have resulted in several paralogs in the Impα gene family, which are usually classified into three conserved clades [58-62]. Both Impα proteins from different organisms, and different Impα paralogs in a single organism exhibit differences in specificity [47, 59, 60, 63-70]; for example, transcription factors Brn2 [66] and STAT1/STAT2 [47] show specificity for Impα5, while RCC1 shows specificity for Impα3/4 family [65] (see also article by Yoneda et al in this issue). However, the conservation of the cNLS-binding sites suggests that differences in binding to this site are unlikely to explain the observed functional differences [41].

It has been suggested that functional cNLSs exhibit upper and lower thresholds in terms of the binding energy for Impα [42]. cNLSs binding Impα weakly are ineffectively imported into the nucleus. The initial rate of protein import has been suggested to be linearly correlated with the cNLS•Impα affinity [53, 71, 72], and is also dependent on the concentration of Impα in the cell [73]. However, sequences binding too tightly, such as Bimax1 and Bimax2 [51] and CBP80 cannot be dissociated from Impα and therefore also cannot function as typical NLSs. Indeed, Bimax1 and Bimax2 can be used as inhibitors of the classical nuclear import pathway, while CBP80 is thought to function through a permanent association with Impα [74].

2.2. Snurportin-1-mediated nuclear import of spliceosomal proteins

The nuclear import of assembled spliceosomal subunits, the uridine-rich small nuclear ribonucleoprotein particles (U snRNPs), involves a variation of the classical nuclear import pathway that uses a distinct adaptor protein called snurportin-1. Snurportin-1 also binds to Impβ1 through an IBB (sIBB) domain [75-77]. Snurportin-1 recognizes an m3G-cap on the spliceosomal RNA [78].

2.3. Nuclear import mediated by direct cargo•importin-β interaction

Some cargo proteins bypass the requirement for an adaptor protein and bind directly to Impβ1. With the exception of the adaptor protein Impα and snurportin-1 IBB domains, there is little similarity in the mode of Impβ1 interaction by these cargo proteins based on the structural data available thus far [79, 80].

Impβ1 plays a central role in the nucleocytoplasmic transport cycle. It interacts with the ccNLS adapter Impα, an assortment of cargo proteins, the Ran protein and nucleoporins. Its ability to bind such a diverse range of proteins can in part be attributed to its large surface area and the inherent flexibility of the solenoid structure. It is composed of 19 tandem HEAT repeats arranged in a superhelix, with each HEAT repeat comprising two helices, the A-helices lining the convex face, and a B-helices lining the concave face.
Most binding partners interact with the B-helices on the concave face, with the exception of nucleoporins that bind on the concave face. This is consistent with the mechanism of nuclear transport; Impβ1 must bind and nucleoporins simultaneously, while Ran binding induces cargo dissociation and is therefore mutually exclusive with cargo binding. The cargo proteins that are recognised by Impβ1 are significantly different from each other. Examples include ribosomal proteins [82], CREB [83], the human immunodeficiency virus (HIV) Rev and Tat [84], SREBP-2 [85], the human T-cell leukemia virus type 1 (HTLV-1) protein Rex [86], PTHrP [87], cyclin B1 [88], Smad3 [89], TRF [90], and SRY [91, 92]. Moreover, the NLSs that confer recognition also vary in both length and number of positively charged residues. The NLS for the Tat protein comprises 9 amino acids, while the NLS for SREBP-2 comprises 120 amino acids. We review the structural basis of recognition by Impβ1 of IBB domains of Impα and snurportin-1, SREBP-2, and PTHrP.

2.3.1. Importin-β recognition of the importin-α IBB domain

The interaction of Impα with Impβ1 occurs through a well-conserved N-terminal αIBB domain, made up of approximately 40 highly conserved amino acid residues. The high-affinity binding by is achieved by Impβ1 wrapping tightly around the αIBB domain, covering over 40% of its surface area. The interaction interface spans HEAT repeats 7-19 (Fig. 2). There are two main regions that mediate interaction: HEAT repeats 7-11, which bind the N-terminal region of the αIBB domain (α11-α23), and HEAT repeats 12-19, which bind the C-terminal α-helical region of the αIBB domain (α24-α51) (Fig. 4) [81].

2.3.2. Importin-β recognition of the snurportin-1 IBB domain

The mechanism by which Impβ1 binds the snurportin-1 IBB (sIBB) domain closely resembles the Impβ•αIBB complex (Fig. 2). The overall conformation of Impβ1 is similar to the closed conformation adopted in the αIBB complex. The sIBB domain (25-65) is bound by Impβ1 in two regions; HEAT repeats 7-11 interact with the N-terminus of sIBB (residues 25-40), while HEAT repeats 12-19 interact with the sIBB helix (residues 41-65). Both regions are critical for binding. The second binding region also exhibits a network of electrostatic interactions, albeit with fewer interactions than observed αIBB helix (12 contacts for the sIBB; 20 for the αIBB).

2.3.3. Importin-β recognition of PTHrP

Parathyroid hormone-related protein (PTHrP) is a peptide hormone with endocrine, autocrine/paracrine and intracrine actions, and is expressed in a range of tumors [93]. Its action relies on its ability to localize to the nucleus, and is recognised directly by Impβ. Nuclear import assays accordingly show Impβ1 and Ran can localise PTHrP to the nucleus in vitro in the absence of Impα [87]. Impβ1 binds PTHrP on its concave face through an extended region spanning HEAT repeats 2-11, with PTHrP running parallel with the superhelical axis of Impβ (Fig. 2). There are three distinct binding surfaces that comprise the determinants of Impβ•PTHrP binding; HEAT repeats 2-7 bind the N-terminus of the PTHrP NLS (67-79), which adopts a β-strand like conformation; HEAT repeats 8-10 bind the central moiety of the PTHrP-
NLS (80-86) structured as a rigid arch; and HEAT repeats 8-11 bind to the C-terminal region of PTHrP (87-93), structured as an extended strand.

Despite the partial overlap of binding sites of PTHrP (HEAT repeats 2-11) with the IBB domains (HEAT repeats 7-19), the architecture and mechanism of binding is significantly different between these cargo molecules.

2.3.4. Importin-β recognition of SREBP-2

The sterol regulatory element binding protein 2 (SREBP-2) is a transcription factor that belongs to the basic-helix-loop-helix (HLH) class of transcription factors and binds to sterol regulatory element DNA sequences to regulate sterol biosynthesis. As the concentrations of cellular sterols levels become low, SREBP-2 is cleaved to release a water soluble N-terminal domain that is targeted to the nucleus by Impβ, where it can up-regulate the expression of enzymes involved in sterol biosynthesis [94].

The structure of Impβ•SREBP-2 HLH domain complex (343-403) reveals that to accommodate binding of SREBP-2, Impβ1 adopts a more open conformation than observed in complex with αIBB and sIBB domains and PTHrP (Fig. 2). The mode of binding is also significantly different, reliant on HEAT repeats 7-17. The SREBP-2 dimer is bound to Impβ1 in an orientation perpendicular to the superhelix, whereas the IBB domains bind in a parallel orientation. Two long helices in HEAT repeats 7 and 17 bind the SREBP-2 dimer and adopt a more twisted and open conformation, acting like chopsticks to bind the SREBP-2 dimer.

In contrast to the electrostatic interactions that dominate the IBB and PTHrP interactions with Impβ, binding of SREBP-2 involves a series of hydrophobic interactions [87].

2.3.5. Similarities and differences in importin-β cargo recognition

In contrast to Impα:cNLS recognition, Impβ1 does not appear to bind its cargo through a conserved mechanism. The IBB domain are bound through a large network of electrostatic interactions, PTHrP binds through electrostatic interactions distinctly different to those involved in binding the IBB domains, while hydrophobic interactions dominate the Impβ•SREBP-2 interaction. One unifying theme that does emerge is that considerable variation exists in the overall structure and helical pitch of the HEAT repeats of Impβ1 upon cargo recognition. This not only observed between different partners, but also within different molecules present in multiple copies within an asymmetric unit or in different crystal forms [95]. For example, the two Impβ1 molecules in the asymmetric unit of the sIBB domain structure vary by up to 20 Å between HEAT repeats 12-19 [76]. This flexibility appears to be crucial for Impβ1 to bind a wide range of cargo as well as accessory proteins important in nuclear protein import.

2.4. Karyopherin-β2 pathway and PY-NLSs

With over 20 proteins involved in mRNA processing identified to be imported by the β-karyopherin member karyopherin-β2 (Kapβ2/transportin-1), a common NLS could not be identified through sequence comparisons. The determination of the structural basis of binding of one of the cargos, complemented by interaction studies of a series of mutants, has allowed the definition of the corresponding NLS termed PY-NLS [96]. Follow up studies refined the rules of NLS recognition [97-99]. Structurally, the NLSs converge at
three sites: an N-terminal motif, a central arginine residue, and the C-terminal PY sequence. Measurements of binding energy of alanine mutants shows a different distribution in different PY-NLSs. Three rules emerge about these NLSs: (1) they are structurally disordered in free cargos; (2) they have an overall basic character; and (3) they have a central hydrophobic or basic motif followed by a C-terminal consensus sequence R/H/KX_{2-5}PY (Fig. 3). The rules led to the identification of over 100 human proteins with PY-NLS, with several experimentally confirmed [96, 100, 101], as well as the design of a high-affinity peptide that can be used as an inhibitor of Kapβ2 pathway [97].

2.5. Other nuclear import pathways

Ten β-karyopherins have been shown to import cargos into the nucleus in humans [5], yet only a few cargos have been identified for most of those, with the exception of Impβ1 and Kapβ2. Because each of these carriers imports a distinct set of cargos, they must be recognizing a different targeting signal. At present however, the definition of an NLS common to different cargos used by a single carrier has only been possible for the classical and Kapβ2-mediated pathways. Many nuclear proteins therefore do not contain a recognizable NLS, and many contain targeting signals that do not resemble the known NLSs [102].

A limited set of cargos is currently known for the β-karyopherin family member importin-13 (Imp13), which is thought to function both in import and export. The known cargos including two components of the exon junction complex, Mago and Y14. The crystal structure of Imp13•Mago•Y14 complex reveals that the transport factor forms a ring around the heterodimeric cargo using an evolutionarily conserved surface, and excludes a cytoplasmic Mago•Y14-binding partner PYM through steric hindrance [103].

Several alternatives to the conventional β-karyopherin-dependent nuclear import pathways exist, including variations that involve the cellular cytoskeletal system to enhance the conventional pathways in cases where rapid transport is critical to function, and pathways using unrelated transporters [104]. Many viral proteins as well as cellular regulatory proteins exploit the microtubular network to increase the efficiency of import, some engaging the motor protein dynein [105]. NF-κB instead uses the actin cytoskeleton [106]. The calcium-binding protein calmodulin has been shown to function as an import factor in a pathway dependent on calcium but independent of β-karyopherins, GTP and Ran, for a range of cargos such as the high-mobility group box and the HLH families of transcription factors [107, 108]. Other proteins enter the nucleus independent of carrier molecules, for example by by binding to nucleoporins themselves or by diffusion followed by binding to nuclear components. Less characterized are intriguing pathways involving lectins as import factors for glycosylated proteins, or involving the disruption of nuclear envelope by viruses during infection. Some proteins appear to use multiple mechanism of nuclear import, perhaps to localize to the nucleus even when conventional pathways are inhibited.

3. Bioinformatic resources of nuclear localization data and for prediction of nuclear localization
Understanding the molecular basis of specificity of nuclear import pathways can help identifying new proteins that function in the nucleus, and which contain a specific targeting sequence and use a specific pathway to enter the nucleus. Diverse types of data are available that are relevant to this task, including subcellular localization data and protein-protein interaction data. A number of bioinformatic resources and tools specific to this task exist. However, none of the resources available currently can answer all the three questions: (1) is the protein nuclear; (2) what pathway does it use to enter the nucleus; and (3) what is the targeting signal? Significant scope exists to improve the tools and integrate disparate information that different resources can provide. We review the currently available resources and tools, and try to establish how (and how well) nuclear localization of proteins can be predicted, to allow biologists to make informed decisions about what tools to utilize and what confidence to put in their output.

3.1. Nuclear localization data resources

Nuclear localization data resources may be classified into two broad categories: (a) comprehensive protein sequence databases that include both experimental and inferred localization metadata, and (b) specialist resources that report on details relevant to nuclear localization. We believe that important insights can be tapped by integrating the qualitatively disparate information contained in these datasets. While it is beyond the scope of this review to discuss the reliability of different types of data, the researcher mining the data needs to be aware that all data come with an inherent level of (un)reliability, in particular data obtained by high-throughput and large-scale approaches.

3.1.1. Comprehensive protein databases with nuclear localization-specific annotations

Many localization datasets encompass multiple organelles including the nucleus. For instance, UniProt [109] and LOCATE [110] both assign broad localization properties to a large number of proteins. UniProt includes proteins from many species and annotations are of varying quality (ranging from hypothetical to experimentally confirmed). An entry for a nuclear protein may contain several relevant sections, including Gene Ontology (GO) terms for the protein’s cellular component, and the amino acid segment that corresponds to the localization signal.

LOCATE is limited to mouse and human proteins, but refers directly to experimental localization evidence if available, including supporting microscopy data. [More about the coverage of LOCATE.] A similar effort for *S. cerevisiae* based on GFP-fusion protein analysis is reported in [111] (http://yeastgfp.yeastgenome.org). The authors show that 1534 yeast proteins (out of ~6,275) are imported into the nucleus (or associate with nucleoli).

There is a growing number of databases that focus on nuclear proteins, and some distinguish between intra-nuclear compartments, e.g. the Nuclear Protein Database [112] (http://npd.hgu.mrc.ac.uk). [More on the coverage of NPD.] NucProt is a focused, mouse-specific reference set that only contains nuclear proteins, of which 2568 are experimentally validated [113]. [More on the coverage of NucProt.] NucProt is useful to illustrate the full nuclear proteome of a single species, but does not reveal the specific means of protein translocation.
3.1.2. Nuclear localization signal resources

Several resources document the variety of NLSs. A large scale screening study of random peptide libraries using mRNA display for yeast cNLS-like sequences that bind to yeast, plant or human Impαs suggested 6 sub-classes of cNLSs: five variants of the monopartite NLS and the a bipartite NLS class. [More on the coverage and access of cNLS data. Maybe include logos of Kosugi’s 6 classes—using their occurrences in the yeast nuclear proteome, Ahmed?]

PY-NLS [96]. [More on this resource in terms of specificity, coverage and data access.]

Non-canonical localization signals are scattered throughout the literature. NLSdb (http://cubic.bioc.columbia.edu/db/NLSdb) contains a set of 114 experimentally validated nuclear import signals directly sourced from the literature [114]. NLSdb does not explicitly distinguish between different pathways or carrier-binding specificities, but a mixture of signals corresponding to classical and other known or yet uncharacterized pathways. The set was artificially extended by in silico mutagenesis to contain 308 signals. This increased the coverage to 43% of known nuclear proteins, while not including any non-nuclear proteins.

This review focused on nuclear import, but it is worth noting that there are more limited resources available for nuclear export signals. NESbase (http://www.cbs.dtu.dk/databases/NESbase) documents 80 specific examples of nuclear export signals [115]. Kosugi et al. [116] report on a “classical” CRM1-dependent export signal library consisting of 101 distinct sequences, further divided into 6 patterns based on hydrophobic amino acid spacing.

3.1.3. Other data resources relevant to nuclear localization

The Protein Data Bank [117] contains experimentally determined protein structures including structures relevant to understanding the structural basis of nuclear cargo protein•carrier recognition. Numerous molecular interaction databases are useful as a source of data on the possible interaction partners for cargo proteins.

3.2. Computational models for prediction of nuclear localization

We can understand most models that attempt to predict nuclear localization by considering two extremes of prediction approaches: models that are based on (a) establishing the similarity of a query protein with already known nuclear proteins (we refer to them as “similarity-based models”), or (b) models based on detecting targeting signals and other properties pertinent to the nuclear import machinery (we refer to them as “NLS-based models”).

The models can also be grouped with regard to the scope of protein destinations they predict: models that only distinguish between nuclear localization and “everything else”, and models that recognize multiple compartments including nucleus (and, for example, mitochondrion, peroxisome, cell membrane). This review is primarily concerned with nuclear import, thus we largely disregard the rich output provided by models that take the multi-compartmental and perhaps more challenging route, unless it clearly helps designate proteins as nuclear.
It is worth noting that most predictors are “trained” on a specific dataset. To provide a realistic estimate of their accuracy, separate “test” data are presented and predictions are compared with known localizations. Caution is required when interpreting accuracy, because test examples may overlap with training data and may be poor representatives of the proteome as a whole. We will discuss these issues in more detail below.

### 3.2.1. Similarity-based models

Most current predictors rely primarily on detecting homology or sharing of generic domains: they essentially measure similarity between a novel protein and proteins with known nuclear (or non-nuclear) designation. To maximize their ability to recover remote homology, the method for assessing this similarity can be based on a variety of techniques including a variety of machine learning methods (e.g. neural networks and support-vector machines) where designated proteins are supplied during a training phase, often as encodings of the amino acid sequence data. A sample of such predictors includes SubLoc [118], ESLpred [119], pSLIP [120], CELLO [121], and Nucleo [122]. Arguably, machine learning techniques can find short targeting signals because of their prevalence in training data but, unless they are specifically designed to capture these, and unless they can explain their predictions in terms of targeting mechanisms, we designate them as primarily homology-based. We acknowledge that some models indeed incorporate information of targeting signals, but very few offer explanations of their predictions that explicitly make reference to features specific to processes of nuclear import.

### 3.2.2. NLS-based models

Direct detection of nuclear localization signals in protein sequence data is an unreliable predictor of nuclear import for at least two reasons. Firstly, NLSs are short and poorly defined; they tend to match by chance many sequences that are non-nuclear. Secondly, NLS definitions are incomplete; known patterns do not cover all nuclear proteins or may not match perfectly to functional signals. These observations, together with the paucity of other predictive features such as the detailed structure of the cargo or the interaction with import machinery, have so far prohibited a wide exploitation in predictors. However, two recent reports deserve a mention.

cNLS Mapper incorporates an additive scoring approach, based on peptide library data obtained by the research group [57] (http://nls-mapper.iab.keio.ac.jp). As the name suggests, the tool is intended to predict cNLSs only. [More on what it does, how it works, and how it compares with PSORT II and PredictNLS in terms of sensitivity.]

NLStradamus [123] (http://www.moseslab.csb.utoronto.ca/NLStradamus) uses a probabilistic model to accommodate subtle variations to classical and non-canonical NLSs. The authors collected 60 NLSs associated with different pathways in the literature. The predictor identifies the NLS, irrespective of the type, in the sequence of the putative cargo. [More on what it does, how it works, and how it compares with PredictNLS (only based on positive data (sensitivity) and NLS location, NOT import).]

[Other predictors that fall into this category:

- PredictNLS [124] (http://www.rostlab.org/services/predictNLS/),
- WoLF PSORT [125] (W585-7; http://wolfpsort.org/),]
• NucPred [126](http://www.sbc.su.se/~maccallr/nucpred/).

Any others?]

3.2.3. Predicting nuclear localization and signals

Models of nuclear translocation often aim to answer questions that would be difficult or time-consuming to answer experimentally. Primarily, we are interested in establishing for a protein whether it is imported into the nucleus. If so, we may ask what pathway and what properties of the cargo are responsible for the translocation. We will compare the available tools to enable researchers to make informed decisions about what computational services to use to answer these questions. Specifically, we ask

1. How well can we predict localization with current tools?
2. Can we predict how a protein is imported?
3. Can we predict what signals are employed?

Most predictors mentioned in this review will give an output that directly scores the tendency of the protein to be imported. In some cases the prediction requires further assessment to answer question 1. For instance, cNLS Mapper predicts the presence of a classical NLS, which is merely indicative of one import pathway. On the other hand, predictors that abstract away from pathway-specific properties are amenable to answer question 1, but since they do not explicitly recognize and provide evidence for NLSs, they are unable to answer questions 2 and 3.

In the following section, we will illustrate the expected accuracy by which the questions can be addressed with the tools most suited to the task.

Studies presenting new models usually offer an assessment of their prediction accuracy in relation to alternative predictors. However, it is not unusual to find that the accuracy varies substantially between different comparisons (that is, test datasets). Hawkins et al. [122] presented a sobering evaluation of seven publicly available nuclear predictors on a test data set, carefully chosen so that it would not overlap with the training data of the tested predictors. The accuracies reported originally gave an average Matthews’ correlation coefficient of 0.71, while the accuracy on the independent data set fell to a low 0.26 (on average; 1.0 is perfect, 0.0 is chance prediction). At the best accuracy, sensitivities ranged from 0.27 up to 0.76, and specificities from 0.53 to 0.93. Best overall was Nucleo [122].

Sprenger et al. [127] similarly performed an evaluation of the overall prediction accuracy of several subcellular localization predictors on basis of proteins in the LOCATE database of mouse proteins. They noted a better accuracy for the nuclear compartment than for many other compartments, with sensitivities ranging from 0.55 to 0.87, and specificities ranging from 0.63 to 0.91. Best overall predictor for nuclear localization was Proteome Analyst [128], though it provides a prediction only when confident (in absolute numbers the authors deemed Proteome Analyst to have the same accuracy as other predictors).

3.2.4. Evaluation of models on mouse and yeast proteomes
To reduce bias in test data selection procedures we set out to construct two proteome-wide data sets, one for mouse and one for yeast, to objectively compare the prediction models listed in this review. Any redundancy in such data sets would simply reflect the true species-specific distribution of sequence sharing. Also, this would enable us to investigate the utility of NLS-based models. Armed with more precise details of effective localization signals, such models are expected to possess greater sensitivity and specificity. We use an NLS-based model to estimate the proteome-wide prevalence of the properties they predict. UNCLEAR

The mouse set is based on Nucprot, supplemented with proteins with their nuclear designation supported by at least two alternative sources, including Uniprot, HPRD (via orthology), and NPD. [More details about the data set.] In total, 3xxx proteins are confidently assigned to the “positive” class. We carefully assembled a set of 3xxx proteins assigned to the “negative” class, by identifying Uniprot proteins with evidence of a specific, non-nuclear localization. We ensured that either class had redundancy exceeding xx% sequence similarity.

The yeast set is directly taken from the yeast-GFP set [111]. This set consists of xxxx “positive” and xxxx “negative” proteins, all with firm experimental evidence of nuclear or non-nuclear localization evidence.

Q1: Use cNLS Mapper and NLStradamus on our data sets to see how “NLS-aware” models are able to deal with the prediction question 1 (see Section 3.2.3). Use Nucleo to relate to the rest of the predictors in the Hawkins study.

Q2: difficult since the NLS based models either only look at one NLS class (cNLS Mapper only looks for classical) or don’t distinguish between them (NLStradamus, PredictNLS lump all signals together), hence making it hard to recover what pathway is used.

3.2.5. Current status of bioinformatic resources on nuclear localization

The value of bioinformatic resources lies in their ability to generate hypotheses on the nuclear localization of a specific protein and its potential import pathway and NLS. To confirm an NLS experimentally, it has to meet four criteria: (1) the sequence is necessary for import, and its deletion or alteration blocks transport; (2) the sequence is sufficient to target an unrelated protein (e.g. GFP) to the nucleus; (3) the cargo protein must interact directly with its import receptor; and (4) disabling of the relevant pathway disrupts import of the protein [17]. Few NLSs have been characterized to such extent. It has to be noted that the results of these experiments may not be unambiguous, due to the complexities of nuclear transport pathways, for example if more than one targeting sequence is present in the protein.
Provide advice to biologist of how to answer Q1-3. Provide advice to computational biologist of the quality of resources, and what to consider when developing new predictive tools.

Both cNLS Mapper and NLStradamus are based on data from yeast, but demonstrate good performance on other species too.

Accuracy-based comments… Including comments on overall import-prediction, NLS-locations etc. Comment on the use of multiple predictors and integrating/combining their predictions. What can be gained?

[More points to come.]

4. Prevalence of different import pathways

Nuclear localization data resources can be used to estimate the prevalence of different nuclear import pathways. The prevalence of cNLSs has been estimated previously in yeast, suggesting that about 60% of nuclear proteins are predicted to use the classical nuclear import pathway [17]. We provide here an update of this estimate. Impα-dependent translocation through the nuclear pore appears to be the most ubiquitously used pathway.

MIKAEL

5. Conclusions

While a number of nuclear import pathways have been identified, the definition of general rules of what constitutes the targeting signal has only been possible for the classical nuclear import pathway (i.e. the cNLS) and the Kapβ2-mediated nuclear import pathway (i.e. the PY-NLS). For other pathways, such rules have not yet been identified for a number of possible reasons, including (i) an insufficient number of cargos is known at the current stage for common features to emerge; (ii) the recognition of cargo by the carrier requires further characterization by structural and interaction analyses; or (iii) no common rules exist and the recognition is specific to specific cargos. For example, few common properties have emerged for cargo proteins recognized directly by Impβ1 despite the availability of structural information on several cargo complexes, and the recognition rules are complex both for cNLSs, and even more so for not only for PY-NLSs. While the identification of rules for NLSs has allowed novel identification of these signals in a number of proteins through sequence comparison-based methods, there is opportunity to develop more powerful bioinformatic search tools that better capture the complex rules of these signals and that will be able to answer the three questions about nuclear localization of proteins: (1) is the protein nuclear; (2) what pathway does it use for nuclear import; and (3) what is the signal that targets it to the nucleus? Such tools have a potential to facilitate and shortcut the otherwise laborious experimental approaches to answer these questions. The computational tools available currently neither answer all the three questions, nor answer many of them with acceptable accuracy. The key to progress is to integrate disparate types of information, from structural through
sequence information to high-throughput data on nuclear localization and protein-protein interactions.

SUPPLEMENTAL DATA

Supplemental Data include?

Acknowledgements

We thank Murray Stewart, Anita Corbett, Yuh Min Chook, Tim Bailey and the members of the Kobe lab for useful discussions. BK is an Australian Research Council (ARC) Federation Fellow and a National Health and Medical Research Council (NHMRC) Honorary Research Fellow.
References


Table 1.
Structural basis of cNLS binding by importin-α. The sequences are aligned based on the interaction with importin-α. “…” indicates additional sequence not shown here. Italics indicate sequence not present in the structural model.

<table>
<thead>
<tr>
<th>Source protein</th>
<th>Minor cNLS binding site</th>
<th>Linker</th>
<th>Major cNLS binding site</th>
<th>Linker length (P1'-P1)</th>
<th>PDB ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tag *</td>
<td>PKKKRKV</td>
<td></td>
<td>PKKKRKV</td>
<td></td>
<td>1E3L, 1BK6</td>
</tr>
<tr>
<td>αIBB domain</td>
<td></td>
<td></td>
<td>EQLKKRKKV</td>
<td></td>
<td>1IQ5, 1Q1T</td>
</tr>
<tr>
<td>AR</td>
<td></td>
<td></td>
<td>GARKKKKLGG</td>
<td></td>
<td>3BTR</td>
</tr>
<tr>
<td>PLSCR1</td>
<td></td>
<td></td>
<td>GIKSKHW</td>
<td></td>
<td>1Y2A</td>
</tr>
<tr>
<td>c-Myc</td>
<td>KVKL</td>
<td></td>
<td>PAAKRVKL</td>
<td></td>
<td>1EE4</td>
</tr>
<tr>
<td>pepTM</td>
<td>KKKREA</td>
<td></td>
<td>PKKKRKREA</td>
<td></td>
<td>3L3Q</td>
</tr>
<tr>
<td>Nup50</td>
<td>MAKRVAE</td>
<td></td>
<td></td>
<td></td>
<td>2CJM, 2C1T</td>
</tr>
<tr>
<td>Nup2p</td>
<td>MNRKIK</td>
<td></td>
<td></td>
<td></td>
<td>1UN0</td>
</tr>
<tr>
<td>TPX2</td>
<td>KKHK...</td>
<td></td>
<td>VMKIKL</td>
<td>42</td>
<td>3KND</td>
</tr>
<tr>
<td>Npl</td>
<td>AVKRP</td>
<td></td>
<td>AAATKKAQAKKKKKLD</td>
<td>11</td>
<td>1E5E, 1EJY</td>
</tr>
<tr>
<td>Rb</td>
<td>KSAGSNPPKPLKKLR</td>
<td></td>
<td></td>
<td>12</td>
<td>1PJM</td>
</tr>
<tr>
<td>N1N2</td>
<td>RKKRK</td>
<td></td>
<td>TEESPLELKDSDKSSK</td>
<td>13</td>
<td>1P1N</td>
</tr>
<tr>
<td>PB2</td>
<td>KRRK</td>
<td></td>
<td>DSSILTDS</td>
<td>13</td>
<td>2J0Q</td>
</tr>
<tr>
<td>CBP80</td>
<td>SRRRHSDEENDGGQP</td>
<td></td>
<td>KRRKTS</td>
<td>13</td>
<td>3FEY, 3FEX</td>
</tr>
</tbody>
</table>

* Abbreviations: TAg, Simian virus 40 large T-antigen; αIBB domain, importin-β binding domain from importin-α; AR, androgen receptor; PLSCR1, phospholipid scramblase 1; c-Myc, human c-Myc proto-oncopogene; pepTM, optimal oriented peptide-library-derived sequence for mouse importin-α; Nup50, nucleoporin Nup50; Nup2, nucleoporin Nup2p; TPX2, target protein for Xenopus kinesin-like protein 2; Npl, nucleoplasm, Rb, retinoblastoma protein; N1N2, Xenopus laevis phosphoprotein N1N2; PB2, influenza virus polymerase PB2; CBP80, human cap binding protein 80.

Table 2.
Databases?

Table 3.
Prediction tools?
**Figure legends**

**Fig. 1.** Structure of importin-α and the mechanism of cNLS binding. (A) Structure of mouse importin-α (cartoon representation, grey) Arm repeat domain, with the peptide corresponding to the nucleoplasmin cNLS bound. The residues binding to the major site residues are shown in red, the linker region is shown in blue and the residues binding to the minor site are shown in yellow. (B) Cartoon representation of mouse importin-α with all currently available structures of cNLS peptides superimposed. The cNLSs all bind to the inner concave groove of the adaptor protein, however, also differentially explore the protein surface (phosphorylated SV40 T-ag cNLS, PDB ID 1Q1S; PLSCR1 cNLS, PDB ID 1Y2A; AR cNLS 3BTR). The structures of the Nup2p/ Nup50 peptides are also shown, which binds to an additional C-terminal site of importin-α (PDB ID 2C1T and 2C1M respectively), as well as the minor site which has important implications in cNLS/importin-α dissociation in the nucleus. **List all shown, what colours** (C) Cartoon representation of mouse importin-α (grey) in the same orientation as A and D, bound to cargo proteins (in ribbon representation) PB2 C-terminal domain fragment (violet blue) and the Cap binding complex (comprised of CBP80, magenta, and CBP20, yellow). The interacting residues of the cargo proteins are shown in stick representation. The CBP80 protein contains a bipartite cNLS, and no other interactions outside this sequence are seen between the adaptor protein and the Cap binding complex. The PB2 cNLS is also bipartite in nature, and the linker region was not visible in the electron density. Interestingly, lysine 718 binds to the P3' binding pocket, and is found outside of the bipartite sequence of PB2 (738KRX12KKIK735). (D) Surface representation of mouse importin-α shown in two orientations with the degree of conservation of each amino acid shown. The highly conserved NLS binding groove on the inner concave surface of the molecule is clearly visible. **I’ll get a better description off Andrew- Not too sure which organisms he has used etc** (E) Schematic representation of a bipartite cNLS, and the binding site residues of mouse importin-α. The conserved tryptophan and asparagine array interact with the aliphatic portions of the basic residues and cNLS backbone respectively. The acidic residues that participate in hydrogen bonding and electrostatic interactions with the basic cNLS side chains are also shown in red, and are found on the periphery of the binding groove.

**Fig. 2.** Structures of importin-β•cargo complexes. (A) Structures of importin-β•cargo complexes (B) Schematic diagram of importin-β•cargo interaction. **Legend**

**Fig. 3.** Structures of karyopherin-β2•PY-NLS complexes. (A) **Legend**
Fig. 1 to be assembled
I think we should show only cargo complexes (all 4)
Worth showing snurportin-1 IBB?
Fig. 4

Bostjan to do