Importin-α is the nuclear import receptor that recognizes cargo proteins carrying conventional basic monopartite and bipartite nuclear localization sequences (NLSs) and facilitates their transport into the nucleus. Bipartite NLSs contain two clusters of basic residues, connected by linkers of variable lengths. To determine the structural basis of the recognition of diverse bipartite NLSs by mammalian importin-α, we co-crystallized a non-autoinhibited mouse receptor protein with peptides corresponding to the NLSs from human retinoblastoma protein and Xenopus laevis phosphoprotein N1N2, containing diverse sequences and lengths of the linker. We show that the basic clusters interact analogously in both NLSs, but the linker sequences adopt different conformations, whereas both make specific contacts with the receptor. The available data allow us to draw general conclusions about the specificity of NLS binding by importin-α and facilitate an improved definition of the consensus sequence of a conventional basic/bipartite NLS (KRK10–12KKRK) that can be used to identify novel nuclear proteins.

Nucleocytoplasmic transport occurs through nuclear pore complexes, large proteinaceous structures that penetrate the double lipid layer of the nuclear envelope. Most macromolecules require an active, signal-mediated transport process that enables the passage of particles up to 25 nm in diameter (~25 MDa). The best characterized nuclear targeting signals are the conventional nuclear localization sequences (NLSs)\(^1\) that contain one or more clusters of basic amino acids (1). The NLSs fall into two distinct classes termed monopartite NLSs, containing a single cluster of basic amino acids, and bipartite NLSs, comprising two basic clusters separated by a spacer.

Despite the variability, the conventional basic NLSs are recognized by the same receptor protein termed importin or karyopherin, a heterodimer of α and β subunits (for recent reviews, see Refs. 2–4). Importin-α (Impα) contains the NLS-binding site, and importin-β (Impβ) is responsible for the translocation of the importin-substrate complex through the nuclear pore complex. Once inside the nucleus, Ran-GTP binds to Impβ and causes the dissociation of the import complex. Impo becomes autoinhibited, and both importin subunits return to the cytoplasm separately without the import cargo. The directionality of nuclear import is conferred by an asymmetric distribution of the GTP- and GDP-bound forms of Ran between the cytoplasm and the nucleus. This distribution is in turn controlled by various Ran-binding regulatory proteins.

Impα consists of two structural and functional domains, a short basic N-terminal Impβ-binding domain (5–7) and a large NLS-binding domain built of armadillo (Arm) repeats (8). The structural basis of monopartite and bipartite NLS recognition by Impα has been studied crystallographically in yeast and mouse Impo proteins (9–11). The two basic clusters of the bipartite NLSs bind to two separate binding sites on Impo, involving Arm repeats 1–4 and 4–8, respectively. Monopartite NLSs can bind in both sites but primarily use the binding site corresponding to the C-terminal basic cluster of the bipartite NLSs, referred to as the major site (9, 11).

The structure of full-length Impo indicated that the major NLS-binding site is occupied by residues 44–54 from the N-terminal region of the protein (Impβ-binding domain) that resembles an NLS (12); Impo is therefore autoinhibited in the absence of Impβ. This observation is supported by the measurements of higher NLS binding affinity by Impαβ as compared with Impo alone (13–22). The significance of the autoinhibitory mechanism has been confirmed by in vitro studies (23).

In this study, we present the crystal structures of peptides corresponding to the bipartite NLSs from human retinoblastoma protein (RB) and Xenopus laevis chromatin assembly factor N1N2 bound to mouse Impo. These NLSs were chosen to

---

\(^*\) This work was supported by the Australian Research Council (to B.K.). The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

This article is dedicated to Alec Hodel.

The atomic coordinates and structure factors (code 1PJM and 1PJP) have been deposited in the Protein Data Bank, Research Collaboratory for Structural Bioinformatics, Rutgers University, New Brunswick, NJ (http://www.rcsb.org/).

†† Supported by the Fundação de Amparo à Pesquisa do Estado de São Paulo, Brazil.

§§ A National Health and Medical Research Council (NHMRC) Senior Research Fellow.

¶¶ Formerly a Wellcome Senior Research Fellow in Medical Science in Australia and currently an NHMRC Senior Research Fellow. To whom correspondence should be addressed. Tel.: 61-7-3365-2132; Fax: 61-7-3365-4699; E-mail: b.kobe@mailbox.uq.edu.au.

\(^1\) The abbreviations used are: NLS, nuclear localization sequence;
Diffraction data statistics

|          | RB peptide | N1N2 peptide |
|----------|------------|--------------|
| a        | 78.0       | 78.9         |
| b        | 89.2       | 89.7         |
| c        | 101.0      | 101.0        |
| Resolution (Å) | 99–2.5 (2.59–2.5) | 99–2.5 (2.59–2.5) |
| Observations | 249,870     | 211,968      |
| Unique reflections | 25,382      | 25,538       |
| Completeness (%) | 96.0 (98.3) | 95.5 (94.6)  |
| Rmerge (%) | 7.3 (40.3)  | 8.2 (51.5)   |
| AverageIo(I) | 19.0 (2.8)  | 15.7 (1.8)   |

Refinement statistics

|          | RB peptide | N1N2 peptide |
|----------|------------|--------------|
| Resolution (Å) | 30–2.5 (2.66–2.5) | 30–2.5 (2.66–2.5) |
| Number of reflections (F>0) | 24,323 (4,075) | 24,340 (4,210) |
| Completeness (%) | 96.0 (98.3) | 95.5 (94.5) |
| Rcryst (%) | 20.4 (26.3) | 20.9 (30.2) |
| Rmerge (%) | 23.9 (29.2) | 23.3 (32.6) |
| Number of non-hydrogen atoms | 3,261 | 3,244 |
| Protein | 146 | 159 |
| Peptide | 173 | 123 |
| Water | 46.2 | 47.2 |
| Mean B-factor (Å²) | 0.006 | 0.007 |
| r.m.s deviations from ideal values | 1.2 | 1.4 |
| Bond lengths (Å) | 93.8 (0.3) | 91.9 (0.3) |
| Ramachandran plot | Calculated with the program CNS (27). |
| Coordinate error (Å) | Calculated with the program PROCHECK (30). |
| Luzzati plot (cross-validated Luzzati plot) | 0.28 (0.35) | 0.29 (0.37) |
| SIGMAA (cross-validated SIGMAA) | 0.24 (0.29) | 0.25 (0.29) |
| Water | 173 | 123 |
| Peptide | 146 | 159 |

TABLE I
Structure determination

|          | RB peptide | N1N2 peptide |
|----------|------------|--------------|
| Resolution (Å) | 30–2.5 (2.66–2.5) | 30–2.5 (2.66–2.5) |
| Number of reflections (F>0) | 24,323 (4,075) | 24,340 (4,210) |
| Completeness (%) | 96.0 (98.3) | 95.5 (94.5) |
| Rcryst (%) | 20.4 (26.3) | 20.9 (30.2) |
| Rmerge (%) | 23.9 (29.2) | 23.3 (32.6) |
| Number of non-hydrogen atoms | 3,261 | 3,244 |
| Protein | 146 | 159 |
| Peptide | 173 | 123 |

Calculated with the program CNS (27).

Calculated with the program PROCHECK (30).

Numbers in parentheses are for the highest resolution shell.

Rmerge = 2Σhkl|Iobs(hkl) − <Ihkl>|/Σhkl|Iobs(hkl)|, where <Ihkl> is the intensity of an individual measurement of the reflection with Miller indices h, k and l, and |Iobs(hkl)| is the mean intensity of that reflection. Calculated for F > 2σ(F).

Rcryst = 1/|Σhkl, i|F(calc)hkl, i| − |F(obs)hkl, i|/Σhkl, i|F(calc)hkl, i|, where |F(calc)hkl| and |F(obs)hkl| are the observed and calculated structure factor amplitudes, respectively.

Rfree is equivalent to Rcryst but calculated with reflections (5%) omitted from the refinement process.

Calculated with the program CNS (27).

Calculated with the program PROCHECK (30).

EXPERIMENTAL PROCEDURES

Peptide Synthesis—The peptides CGKRSAEGSNPPKPLKKGY (RB peptide) and CGKRKKTEERESPLKDKAKKSKGY (N1N2 peptide) were synthesized using the Applied Biosystems 433A peptide synthesizer, purified by cation exchange chromatography followed by reverse phase chromatography, and analyzed by quantitative amino acid analysis using a Beckman 6300 amino acid analyzer and electrospray mass spectrometry (Sciex API 111, PerkinElmer Life Sciences) (24). The peptides RB and N1N2 correspond to the NLSs of human retinoblastoma protein, residues 861–877, and X. laevis N1N2 protein (N1N2), residues 535–555, respectively, with two heterologous residues added at each terminus.

Protein Expression, Purification, and Crystallization—N-terminal truncated mouse importin-α (α2 isoform (25)) lacking 69 N-terminal residues (m-Impα) was expressed recombinantly in Escherichia coli as a fusion protein containing a hexa-histidine tag (11). For crystallization, m-Impα was concentrated to 18.8 mg/ml (in 20 mM Tris-HCl (pH 8.0), 100 mM NaCl, and 10 mM dithiothreitol) using a Centricon-30 (Millipore) and stored at −20 °C. Crystallization conditions successful for other peptide complexes (11) as a starting point. The crystals of both complexes (rod-shaped, 0.5 × 0.2 × 0.1 mm for the RB peptide complex and 0.4 × 0.1 × 0.07 mm for the N1N2 peptide complex) were obtained using co-crystallization by combining 1 μl of protein solution, 0.7 μl of peptide solution (1.7 mg/ml with peptide/protein ratio 3.5), and 1 μl of reservoir solution and suspended over 0.5 ml of reservoir solution containing 0.6 M sodium citrate (pH 6.0) and 0.1 M sodium dithiothreitol.

Diffraction Data Collection—The crystals exhibit orthorhombic symmetry (space group P2₁2₁2₁; Table I). Diffraction data were collected from single crystals transiently soaked in a solution analogous to the reservoir solution but supplemented with 23% glycerol and flash-cooled at 100 K in a nitrogen stream (Oxford Cryosystems), using a MAR-Research image plate detector (plate diameter, 345 mm) and CuKα radiation from a Rigaku RU-200 rotating anode generator. Data were autoindexed and processed with the HKL suite (26) (Table I).

Structure Determination and Refinement—The crystals of the peptide complexes were highly isomorphous with the crystals of full-length Impα (12); therefore, the structure of mouse Impα (Protein Data Bank number 1IAL) with N-terminal residues omitted was used as a starting model for crystallographic refinement. Electron density maps were inspected for the presence of the peptide after rigid body refinement using the program CNS (27) (RB peptide: m-Impα complex, Rcryst = 30.5%, Rfree = 32.4%, 6–4-Å resolution; N1N2 peptide: m-Impα complex, Rcryst = 30.8%, Rfree = 35.1%, 6–4-Å resolution). Table I provides the explanation of R-factors. Electron density maps calculated with coefficients 3 Fobs − 2 Fcalc and simulated annealing omit maps (Fig. 1) calculated with analogous coefficients were generally used. The model was improved, as judged by the free R-factor (28), through rounds of crystallographic refinement (positional and restrained isotropic individual B-factor refinement with an overall anisotropic temperature factor and bulk solvent correction) and manual rebuilding (program O (29)). Solvent molecules were added with the program CNS (27). Ass390 is an outlier in the Ramachandran plot as also observed in all other structures of mouse Impα (11, 12). Pro242 is a cis-proline. The final models comprise 427 Impα residues (residues 71–497), 20 peptide residues, and 173 water molecules for RB...
peptide-Impα complex, and 426 Impα residues (residues 72–496), 21 peptide residues, and 123 water molecules for N1N2 peptide-Impα complex (Table I). The coordinates have been deposited in the Protein Data Bank (Protein Data Bank numbers 1PJM and 1PNJ for the RB and N1N2 peptide complexes, respectively).

Structure Analysis—The quality of the models was assessed with the program PROCHECK (39). The contacts were analyzed with the program CONTACT, and the buried surface areas were calculated using the program CNS (27).

Bioinformatic Analysis—We used the consensus sequence KRKK to search for NLS-containing proteins using the Quick Matrix option of PredictNLS (33); this method detected an NLS in nine proteins, some of which did not belong to the conventional basic/bipartite group.

RESULTS AND DISCUSSION

Structure Determination—The RB and N1N2 NLS peptides were co-crystallized with an N-terminal truncated mouse Impα lacking residues 1–69 (m-Impα); residues 1–69 are responsible for autoinhibition. The co-crystals with both peptides grew in similar conditions and isomorphously to other mouse Impα crystals (11, 12). Electron density maps based on the Impα model, following rigid body refinement, clearly showed electron density corresponding to the peptides (Fig. 1). The structures were refined at 2.5-Å resolution for both complexes (Table I).

Residues 859–878 of the peptide RB (residue 879 had no interpretable electron density) and residues 533–555 of the peptide N1N2 (residues 533–534 and 556 had no interpretable electron density) could unambiguously be identified in the electron density maps (Fig. 1). The side chains of Glu543 and Lys549 of N1N2 were poorly ordered; therefore, these residues were modeled as alanines.

Structure of Importin-α in the Complexes—Impα forms a single elongated domain built from 10 Arm structural repeats, each containing three α helices (H1, H2, and H3) connected by loops (Fig. 2). The structure of Impα in the complexes is comparable with the crystal structure of the full-length Impα. (r.m.s. deviations of Ca atoms of Impα residues 72–496 are 0.22 Å between the RB and N1N2 peptide-m-Impα complexes; 0.35 and 0.31 Å between full-length Impα and the RB and N1N2 peptide-m-Impα complexes, respectively; and 0.34 and 0.30 Å between nucleoplasmin peptide-m-Impα and the RB and N1N2 peptide-m-Impα complexes, respectively.)

Binding of the NLS Peptides to Importin-α—The peptides bind in an extended conformation with the chain running antiparallel to the direction of the Arm repeat superhelix (Fig. 2). The base of the groove that contains the binding sites is formed mainly by the H3 helices of the Arm repeats, which carry some residues conserved among the repeats, including the tryptophans and asparagines at the third and fourth turns in H3 helices of the Arm repeats, respectively (10, 11).

The two basic clusters of the RB and N1N2 peptides bind to two separate well defined binding sites on the surface of the m-Impα molecule, referred to as the minor and major sites (Fig. 2). The minor site specifically binds to the N-terminal basic cluster KR, and the larger, C-terminal basic cluster binds to the major site. The electron density is present for 20 peptide residues in the RB peptide (average B-factor, 63.3 Å²) and for 21 peptide residues of the N1N2 peptide (average B-factor, 72.1 Å²) (Fig. 1). In both complexes, the linker sequences connecting the major and minor sites (residues 863–873 and 539–550 for RB and N1N2, respectively) have B-factors above the average number for the entire peptides (71.1 and 82.8 Å² for RB and N1N2, respectively). By contrast, the residues bound to the major sites of both complexes have lower B-factors (40.5 and 47.6 Å² for position P2–P5 of RB and N1N2, respectively).
The binding of both basic clusters (positions P₁–P₅ and P₃′–P₅′) is similar in all the available structures, and the major differences occur in the linker regions connecting the basic clusters (Fig. 4). The exceptions are the N- and C-terminal portions of nucleoplasmin NLS bound to m-Impα, where the side chains of Lys165 (position P₁′; N terminus of the peptide) and Lys170 (position P₅; C terminus of the peptide) follow the direction of the main chain of the other peptides. Because all the other peptides contain at least one additional residue at the N and C termini, the conformation of the nucleoplasmin NLS peptide may be an artifact of the short length of the peptide.

The case of nucleoplasmin highlights the importance of residues preceding position P₁′ and following position P₅.

The structures of nucleoplasmin NLS bound to y-Impα and m-Impα are significantly different (r.m.s. deviation of Ca atoms of residues 155–170 is 2.44 Å). In addition to the differences caused by the different lengths of the peptides discussed above, some differences may be explained by structural differences between the two Impα proteins; the yeast structure is slightly more “open” than the mouse structure (12). Most of the differences are found in the region comprising residues 159–165 (there is high structural similarity when only the major (r.m.s. deviation of Ca atoms for positions P₁–P₅ is 0.28 Å) and minor (r.m.s. deviation of Ca atoms for positions P₁′–P₅′ is 0.24 Å) sites are superimposed). With the basic clusters binding most tightly, the different curvatures of the two Impα proteins appear to be compensated in the linker region. Although some linker region residues have different conformations, the main contacts of this region with Impα are comparable in both nucleoplasmin structures. The most important interactions occur for the main chain of conserved residues Arg315, Arg238, and Tyr277 of m-Impα (Arg231, Arg244, and Tyr283 of y-Impα).

The portions of RB and N1N2 peptides bound in the major and minor sites superimpose closely with the nucleoplasmin NLS (the r.m.s. deviations of Ca atoms of positions P₁–P₅ and positions P₁′–P₅′ are 0.34 and 0.42 Å between RB and nucleoplasmin and 0.33 and 0.42 Å between N1N2 and nucleoplasmin, respectively). The structure of the nucleoplasmin linker region is more similar to that of RB than that of N1N2 (r.m.s. deviation of Ca atoms of linker region is 1.70 Å between RB (864–873) and nucleoplasmin (157–165) and 2.20 Å between N1N2 (541–550) and nucleoplasmin (157–165), mainly in the region closest to the major site (residues 162–165 for nucleoplasmin). The region of the linker closest to the major site is structurally conserved best among the three bipartite NLS peptides.

**Binding of Bipartite NLS Linker Regions**—The nucleoplasmin NLS-Impα complexes (10, 11) showed that the main chain of the linker region binds to the conserved Impα residues Arg315 (Arm 6) and Tyr277 (Arm 5) (Arg231 and Tyr283 for y-Impα; these residues interrupt the regularity of Trp-Asn array (11)). The conserved residue Arg238 (Arm 4) (Arg244 for y-Impα) also binds to the main chain of nucleoplasmin NLS. By contrast, the structures of RB and N1N2-m-Impα complexes show the binding of Arg231 and Tyr277 to the side chains of peptides (Arg238 binds to the main chain of the peptides). These observations suggest that residues Arg231, Tyr277, and Arg238 play crucial roles in binding bipartite NLSs; however, these interactions are specific to individual NLSs.

Other important contacts of the peptide linker with m-Impα involve the residues Asn868 and Lys871 for the RB peptide and Lys857 for the N1N2 peptide; all of them are situated close to the major site. This is the region of the linker most structurally conserved among the three bipartite NLS peptides. The electron density maps of RB and N1N2 have a superior quality in this region as compared with the rest of linker. These data reflect the strong interaction of these residues with the protein.

There is 2573 and 2715 Å² of surface area buried between m-Impα and the RB and N1N2 peptide, respectively. All residues of the RB peptide, except residues 865 and 869, and of the N1N2 peptide, except residues 542 and 546, make contacts with m-Impα at distances below 4 Å (Fig. 3). The minor and major site portions of the RB and N1N2 peptides have very similar structures (Fig. 4); after superposition of the equivalent Ca atoms, the r.m.s. deviation of the residues in positions P₁–P₅ is 0.19 Å, and the r.m.s. deviation of the residues in positions P₁′–P₅′ is 0.03 Å. By contrast, they adopt very different conformations in the linker regions between positions P₁′ and P₁; the path of the peptide chain is more linear in the case of RB than N1N2. The bipartite NLS linker sequence of both peptides makes favorable interactions with the H3 helices of armadillo repeats 4–7 of Impα. The most important contacts are limited to residues Asn868 and Lys871 of RB and Lys857 of N1N2.

Among other interactions, the conserved residues Arg315 (Arm 6) and Tyr277 (Arm 5) of Impα that interrupt the regularity of the Trp-Asn array (11) make extensive main chain and side chain contacts with the peptides. These residues also interact with nucleoplasmin NLS in mammalian and yeast Impα (10, 11). However, in that case, the interaction involves only the main chain of the peptide. Other important contacts are made with Arg238 (main chain of RB and N1N2), Ser278 (side chain of N1N2), and Ser234 (side chain of RB).

**Comparison with Other NLS Peptide-Importin-α Complex Structures**—Significantly, the work presented here allows us for the first time to perform a detailed comparison of the structural determinants of binding of a number of NLSs to Impα, to align the NLSs with the binding pockets on Impα with some confidence, and to draw some general conclusions on the specificity of NLS binding (Table II). Despite diverse sequences, the binding of both basic clusters (positions P₁–P₅ and P₃′–P₅′) is similar in all the available structures, and the major differences occur in the linker regions connecting the basic clusters (Fig. 4). The exceptions are the N- and C-terminal portions of nucleoplasmin NLS bound to m-Impα, where the side chains of Lys165 (position P₁′; N terminus of the peptide) and Lys170 (position P₅; C terminus of the peptide) follow the direction of the main chain of the other peptides. Because all the other peptides contain at least one additional residue at the N and C termini, the conformation of the nucleoplasmin NLS peptide may be an artifact of the short length of the peptide.

The case of nucleoplasmin highlights the importance of residues preceding position P₁′ and following position P₅.

The structures of nucleoplasmin NLS bound to y-Impα and m-Impα are significantly different (r.m.s. deviation of Ca atoms of residues 155–170 is 2.44 Å). In addition to the differences caused by the different lengths of the peptides discussed above, some differences may be explained by structural differences between the two Impα proteins; the yeast structure is slightly more “open” than the mouse structure (12). Most of the differences are found in the region comprising residues 159–165 (there is high structural similarity when only the major (r.m.s. deviation of Ca atoms for positions P₁–P₅ is 0.28 Å) and minor (r.m.s. deviation of Ca atoms for positions P₁′–P₅′ is 0.24 Å) sites are superimposed). With the basic clusters binding most tightly, the different curvatures of the two Impα proteins appear to be compensated in the linker region. Although some linker region residues have different conformations, the main contacts of this region with Impα are comparable in both nucleoplasmin structures. The most important interactions occur for the main chain of conserved residues Arg315, Arg238, and Tyr277 of m-Impα (Arg231, Arg244, and Tyr283 of y-Impα).

The portions of RB and N1N2 peptides bound in the major and minor sites superimpose closely with the nucleoplasmin NLS (the r.m.s. deviations of Ca atoms of positions P₁–P₅ and positions P₁′–P₅′ are 0.34 and 0.42 Å between RB and nucleoplasmin and 0.33 and 0.42 Å between N1N2 and nucleoplasmin, respectively). The structure of the nucleoplasmin linker region is more similar to that of RB than that of N1N2 (r.m.s. deviation of Ca atoms of linker region is 1.70 Å between RB (864–873) and nucleoplasmin (157–165) and 2.20 Å between N1N2 (541–550) and nucleoplasmin (157–165), mainly in the region closest to the major site (residues 162–165 for nucleoplasmin). The region of the linker closest to the major site is structurally conserved best among the three bipartite NLS peptides.

**Binding of Bipartite NLS Linker Regions**—The nucleoplasmin NLS-Impα complexes (10, 11) showed that the main chain of the linker region binds to the conserved Impα residues Arg315 (Arm 6) and Tyr277 (Arm 5) (Arg231 and Tyr283 for y-Impα; these residues interrupt the regularity of Trp-Asn array (11)). The conserved residue Arg238 (Arm 4) (Arg244 for y-Impα) also binds to the main chain of nucleoplasmin NLS. By contrast, the structures of RB and N1N2-m-Impα complexes show the binding of Arg231 and Tyr277 to the side chains of peptides (Arg238 binds to the main chain of the peptides). These observations suggest that residues Arg231, Tyr277, and Arg238 play crucial roles in binding bipartite NLSs; however, these interactions are specific to individual NLSs.

Other important contacts of the peptide linker with m-Impα involve the residues Asn868 and Lys871 for the RB peptide and Lys857 for the N1N2 peptide; all of them are situated close to the major site. This is the region of the linker most structurally conserved among the three bipartite NLS peptides. The electron density maps of RB and N1N2 have a superior quality in this region as compared with the rest of linker. These data
suggest that the C-terminal portion of the linker (closest to the major site) plays an important role in binding to Imp/H9251. These data are supported by NLS binding studies (15, 17, 19, 21) and structural analyses of extended simian virus 40 (SV40) large tumor-antigen (T-Ag) peptide complexes. The electron density maps suggest that the RB peptide is better ordered overall, consistent with a larger number of contacts.

Not every linker residue is able to make contacts with the protein. Because of the differences in the lengths of the linkers between RB (11 residues, if we define the linker sequence as the sequence between sites P2\textsubscript{H11032} and P2) and N1N2 (12 residues), the RB peptide binds to m-Imp/H9251 in a more extended conformation than the N1N2. The 11-residue length appears more favorable for binding; 12 residues require some short turns and force the chain farther from Imp/H9251, precluding some favorable interactions. Importantly, this is consistent with the observa-

\[ \text{FIG. 3. Peptide-importin-\alpha interaction. A, schematic diagram of the interactions between the RB peptide and m-Imp. Polar contacts are shown with dashed lines, and hydrophobic contacts are indicated by arcs with radiating spokes. The NLS peptide residues are labeled with R. The water molecules are labeled with S. Carbon, nitrogen, and oxygen atoms are shown in black, white, and gray, respectively. This figure was prepared with the program LIGPLOT (41). B, schematic diagram of the interactions between the N1N2 peptide and m-Imp, shown as in A. The NLS peptide residues are labeled with N.} \]

\[ \text{FIG. 4. Superposition of the RB (green), N1N2 (red), and nucleoplasmin (cyan) (11) NLS peptides bound to m-Imp. The Ca atoms of m-Imp in the three complex structures were used in the superposition. This figure was drawn with the program RIBBONS (40).} \]

\[ \text{M. R. M. Fontes, T. Teh, G. Toth, A. John, I. Pavo, D. A. Jans, and B. Kobe, unpublished results.} \]
tion that incorporation of QPWL in the linker region of nucleoplasmin NLS reduced the efficiency of nuclear import (34).

The affinities for mouse Impα of N1N2 and RB NLSs fused to β-galactosidase have been determined using enzyme-linked immunosorbent assay assays (15, 19). N1N2 binds with a higher affinity than RB to both Impα/β complex (Kd measured as 5.4 nM for N1N2 and 45 nM for RB) and Impα alone (Kd measured as 22 or N1N2 and 180 nM for RB). Therefore, although the linker region appears to interact more favorably in the case of RB, it is most likely the presence of Lys in position P2 that is responsible for the higher affinity of the recognition of N1N2 NLS by Impα (see below).

The Role of Minor Site Binding—Mutagenesis studies with nucleoplasmin (34), RB (15), and N1N2 (19) NLSs revealed that substitutions of P1 and P2 residues to other than Arg or Lys abolished nuclear localization. The binding of Lys-Arg at positions P1 and P2 is observed in all available bipartite NLS peptide-Importin-a complex structures. Similarly, the majority of the bipartite NLSs characterized contain a Lys-Arg sequence in the N-terminal cluster (35). Comparison of the structures here reveals a high structural similarity of the binding of the Arg side chain in the position P2 in all cases. The Arg side chain is situated at the groove created by the tryptophans Trp399 and Trp405 (Trp385 and Trp383 for y-Impα) located in H3 helices of Arm repeats 7 and 8 and also contacts Glu396 and Ser366 (Glu402 and Ser366 in y-Impα); these residues are conserved among known Impα sequences. The interaction with Glu396 appears particularly important because the distance from the Arg side chain is nearly identical in all the structures. The binding of T-Ag at the minor site (11) represents a model for binding of Lys (instead of Arg) at P2′; the peptide 126PKKKRKV132 places Lys128 and Lys129 at positions P1′ and P2′, respectively. This likely occurs because it is more favorable to have some amino acid binding at positions P1′ and P2′, rather than having an Arg at P2′. However, some evidence of staggering in different registers was observed in the crystal structure (11). The Lys at position P2′ of T-Ag contacts Glu396 and Ser366 at approximately the same distance but does not contact Trp399, losing one favorable contact. The nonapartite NLSs from c-Myc binds with Lys-Arg in the P1′–P2′ positions (10).

A Lys side chain at position P1′ makes less favorable contacts than Arg at position P2′, and this position appears less important than the P2′ position for bipartite NLS binding. The P1′ Lys of RB and N1N2 peptides makes interactions with the side chains of the conserved Thr328 and Asn364 and the main chain of Val521. Also close by is Asp293. The P1′ Lys of nucleoplasmin bound to y-Impα also contacts the equivalent of Thr328 and Val521 but not Asn364. The pocket prefers a Lys residue because an Arg side chain is too long to make similar favorable contacts in the P1′ pocket.

Finally, we suggest that the positions preceding the P1′ and following P2′ contribute significantly to the minor site binding. This is consistent with the side chain of the N-terminal Lys126 (position P1′) of the nucleoplasmin peptide preferring to follow the main chain of the other peptides instead of binding in the regular P1′ pocket. Also, the T-Ag peptide 126PKKKRKV132 binds to Impα with two Lys residues in P1′, and P2′ instead of the more favorable Arg at P2′, possibly so that it can place some amino acid in positions P4′ and P5′.

The NLS Consensus—The definition of an NLS consensus based on an analysis of sequences alone is difficult due to the diversity of sequences that can function as nuclear targeting signals. The structural data, the accumulated knowledge on functional NLSs, and mutagenesis studies now allow us to better define the conventional basic/bipartite NLS consensus that involves Impα binding. The ability to recognize NLSs is of significance, as it could help with the functional annotation of new proteins predicted in genomic sequences.

We used an approach to define the NLS consensus similar to that used previously to define optimal substrates for protein kinases (36). The structures of Impα-NLS peptides were analyzed by molecular modeling in conjunction with other available information. This allows us to draw some conclusions about individual positions in the NLS.

One general observation is that the binding pockets for individual side chains are much better defined in the major site than in the minor site. This observation is consistent with the contribution of the N-terminal basic cluster (minor site) in a bipartite NLS of ~4 kcal/mol, which is comparable with the loss of a single Lys side chain in the major basic cluster (21). A large portion of the free energy of binding is contributed by the main chain of the peptide (estimated at ~6.9 kcal/mol), but a functional NLS requires an additional contribution of the side chains of about 4.2 kcal/mol; this contribution by the side chains accounts for the specificity of nuclear import. The best defined pockets include P2, P4, and P5 of the major site. P2 is defined mainly by residues Thr155 and Asp192 in adjacent Arm repeats of Impα and is well suited for binding a Lys side chain. Arg can be accommodated only to a lesser degree, consistent with the loss of 2.7 kcal/mol for the Lys to Arg mutation in the context of T-Ag (21). Side chains such as Thr, Glu, and Pro could be accommodated but would contribute less to binding; there is a substantial electrostatic contribution to binding in this pocket. P4, comprising Glu266 and Asp270 as the major binding determinant, prefers Arg over Lys based on Ala mutagenesis; this is likely because an Arg can reach closer to Glu266 and Asp270. The energetic contribution of the side chains in the P4 site is about two-thirds of the contribution in P2, with a smaller contribution of the electrostatic terms. The modeling suggests that P5, with the primary binding residue Gln81, is best suited for binding a Lys side chain. The electrostatic contribution is very minor, and positive charge is not strictly required; due to size, Glu and Gln would be favored over Arg. The energetic contribution of this site is similar to P3 (21).

The other pockets are less specific. The P4 pocket is relatively large and should favor Arg because this is the only side chain that can reach to the main chain of Impα Arg106 and Glu107 and make favorable interactions. However, hydrophobic side chains such as Pro, Val, Ile, or Leu could also be accommodated relatively well. The discrimination between Arg and Val is only
minor based on the Ala mutagenesis results of T-Ag and c-Myc NLSs, and the contribution of this pocket is about four times smaller than that of P$_2$, P$_1$, with Asp$^{770}$ responsible for specificity, prefers Lys over Arg, whereas side chains such as Gln, Pro, or Gly could easily be tolerated, and the side chain contribution in this pocket is only half that of pocket P$_2$. The pocket preceding P$_1$ has Trp$^{231}$ as the major specificity determinant and also prefers Lys over residues such as Arg, Pro, Ala, Gru, or Gly, but is also not very specific.

The minor site is much less defined than the major site. Only the sites P$_1'$ and P$_2'$ appear to be able of significant discrimination between different peptide side chains, with Lys and Arg favored in those two positions, respectively, as discussed above. The N-terminal basic cluster may play a role in relieving autoinhibition (the minor site is not autoinhibitory).

We conclude that the optimal basic/bipartite consensus sequence for binding Imp$\alpha$ is KRX$_{10-12}$KRRK, with Lys at position P$_2$ (bold and underlined) the most important specificity determinant that also forces the rest of the sequence to bind in one register. The basic residues at positions P$_1$ and P$_3$ (underlined) are also significant (Table II). To facilitate accurate NLS identification in novel sequences, the motif could be represented in a matrix format with probabilities for each of the 20 amino acids defined at each position (31); this should be possible when experimental data are available through a peptide library (37) or equivalent experiment.

Our discussion assumes that there is a linear correlation between Imp$\alpha$ binding and the rate of nuclear import, as the available data suggest (15, 16, 38). However, a functional NLS may display both lower and upper limits in affinity (the NLS available data suggest (15, 16, 38). However, a functional NLS does not have a significant predictive power, and thus is complementary to the approach of defining the conventional basic/bipartite NLS consensus. A comparison of the efficiency of NLS detection in known nuclear proteins between PredictNLS and our consensus-based search using Scansite (31) revealed the latter approach to be superior, although a significant portion of the nuclear proteins does not utilize the Impo/Impβ-dependent import pathway (see “Experimental Procedures”).

Conclusions—The structures of the complexes of mammalian Impo with the peptides RB and NIN2, corresponding to bipartite NLSs, and their comparisons with other available Impo structures, provide new insights into the molecular basis of nuclear import and its regulation. Although the binding of both basic clusters is comparable in all the structures, the linker region connecting the two basic clusters of an NLS makes different but specific contacts with Impo in the cases of all three common linker lengths (10, 11, and 12 residues). The most critical region in the linker for binding is the extreme C-terminal end of the linker sequence. The most important residue binding in the minor site is Arg at position P$_2'$ of the NLS, although the surrounding residues make significant contributions. The integration of the structural information with sequences of characterized NLSs and mutagenesis data allows us to provide an improved consensus sequence for the conventional basic/bipartite NLS, KRX$_{10-12}$KRRK.
Structural Basis for the Specificity of Bipartite Nuclear Localization Sequence Binding by Importin-α

Marcos R. M. Fontes, Trael Teh, David Jans, Ross I. Brinkworth and Bostjan Kobe

J. Biol. Chem. 2003, 278:27981-27987. doi: 10.1074/jbc.M303275200 originally published online April 14, 2003

Access the most updated version of this article at doi: 10.1074/jbc.M303275200

Alerts:
  • When this article is cited
  • When a correction for this article is posted

Click here to choose from all of JBC's e-mail alerts

This article cites 40 references, 13 of which can be accessed free at http://www.jbc.org/content/278/30/27981.full.html#ref-list-1