The nuclear location signal

BY A. E. SMITH1,2, D. KALDERON2†, B. L. ROBERTS1,2, W. H. COLLEDGE2, M. EDGE3, P. GILLET2, A. MARKHAM3,
EVA PAUCHA2‡ AND W. D. RICHARDSON2

1 Integrated Genetics, 31 New York Avenue, Framingham, Massachusetts 01710, U.S.A.
2 Biochemistry Division, National Institute for Medical Research, Mill Hill, London NW7 1AA, U.K.
3 I.C.I. Pharmaceuticals Division, Alderley Park, Cheshire SK10 5TG, U.K.

A short sequence of predominantly basic amino acids Pro-Pro-Lys-Lys-Arg-Lys-Val from SV40 Large T is responsible for the normal nuclear location of the protein. Alteration of Lys-128 to each of six different residues other than Arg renders Large T cytoplasmic, whereas single amino acid changes in the surrounding region impair but do not prevent nuclear accumulation. When transposed to the amino terminus of cytoplasmic Large T species, or Escherichia coli β-galactosidase or of chicken muscle pyruvate kinase, the sequence around Lys-128 of Large T is able to direct the recipient protein to the nucleus. This demonstrates that these amino acids can be sufficient for nuclear location and can act as a nuclear location signal. A computer search of over 2500 proteins reveals that some other nuclear proteins (for example, BK virus Large T, SV40 VP2 and adenovirus 72kDa DNA binding protein) contain very similar basic tracts, but so too do some presumed non-nuclear proteins (for example, poliovirus VP3). We suggest that the related sequence acts as the nuclear location signal in the other nuclear proteins but that the sequence does not function in all cases, perhaps because it is not accessible. A similar, but shorter or less basic sequence, was detected in a number of other nuclear proteins, for example, polyoma virus Large T, SV40 VP1 and several histones. However, such sequences were also found in many other proteins. Perhaps the shorter basic sequences can also act as nuclear location signals, but to be functional they need to be exposed (for example, at the amino terminus of the protein as in SV40 VP1) or to be present in multiple copies.

INTRODUCTION

Until recently, most studies on the structural requirements for the nuclear location of proteins involved the microinjection of proteins and other macromolecules into cells followed by studies of their subcellular distribution (Gurdon 1970; Feldherr

† Present address: Department of Biochemistry, University of California, Berkeley, California 94720, U.S.A.
‡ Department of Neoplastic Disease Mechanisms, Dana-Farber Cancer Institute, 44 Binney Street, Boston, Massachusetts 02181, U.S.A.
Such studies have led to the view that many macromolecules may enter the nucleus by diffusion and those that concentrate there do so by binding to fixed nuclear components. While this may explain the nuclear location of some small proteins, the existence of a specific transport mechanism has also been postulated (De Robertis et al. 1978) to account for the rapid and selective distribution of larger proteins between the nucleus and cytoplasm.

Although more powerful physical methods have recently been used to study nuclear location (Peters 1984), very little is known about the molecular requirements for this process and almost nothing of the postulated transport mechanism. The best characterized system until recently was *Xenopus* oocyte nucleoplasmin which is thought to be transported to the nucleus as a pentamer, under the influence minimally of a single copy of a 15000 Da domain of the protein (Dingwall et al. 1982).

In the course of our studies on Large T, the transforming protein of SV40, we have isolated mutants that are defective for nuclear location (Kalderon et al. 1984a; Kalderon & Smith 1984). The cytoplasmic variants cluster to a very small region of the Large T gene and have enabled us to identify the nuclear location signal of the protein. We have subsequently demonstrated that the same sequence acts as a signal to translocate otherwise cytoplasmic proteins to the nucleus (Kalderon et al. 1984b). Here we review our experimental work to identify the minimal sequence that can act as nuclear location signal. We also report a computer search for similar sequences in other proteins.

**Nuclear location signal of SV40 Large T**

SV40 Large T has a number of biochemical activities, among them the ability to bind to the origin of replication on SV40 DNA (Tooze 1981). A domain on Large T that is required for origin DNA binding has been mapped between residues 83 and 250 on Large T (reviewed in Paucha et al. 1985a). As a part of our studies to ask whether this domain is required for transformation by SV40 Large T, we have introduced extensive mutations into the DNA encoding this region of the protein (Kalderon & Smith 1984).

It transpired that among the mutants we isolated were some that lacked the ability to bind to SV40 origin DNA (Paucha et al. 1985a), as predicted, but also

---

**Description of Plate 1**

*Figure 1.* Subcellular location of SV40 large T and pyruvate kinase variants. Subconfluent Vero (African green monkey) cells were microinjected with plasmid DNA encoding wild-type or mutant SV40 Large T (a)–(c), or fusion proteins containing portions of Large T linked to amino acids 17–529 of pyruvate kinase (PK) (d)–(f). Approximately 18 h later the cells were fixed, and proteins made visible by immunofluorescence microscopy. SV40 Large T was detected by using as first antibody a monoclonal directed against its carboxy-terminal region. (a) Wild-type Large T; (b) d10 mutant Large T (Lys 128 → Thr); (c) d1 mutant Large T (Lys 129 → Met, Lys 131 → Thr). Large T–PK fusion proteins were labelled with a polyclonal antibody raised in rabbits against purified PK. (d) Amino acids 1–136 of d10 mutant Large T (Lys 128 → Thr) fused to PK; (e) amino acids 1–136 of wild-type Large T fused to PK; (f) amino acids 126–132 of wild-type Large T fused to PK.
Figure 1. For description see opposite.
The nuclear location signal

others that had lost the ability to localize to the nucleus (Kalderon et al. 1984a). This could be seen in Rat-1 cells transformed by the mutants, but it was most rapidly and most readily observed by microinjecting about 200 copies of the mutated plasmid DNAs into the nucleus of Vero cells (Richardson & Westphal 1981) and 12-16 h later visualizing the protein expressed by immunofluorescence. By using this assay, wild-type DNA gave very intense nuclear fluorescence with dark nucleoli and no cytoplasmic staining (figure 1a). By contrast, when Lys-128 was converted to Thr (d10), the Large T appeared to be located almost exclusively in the cytoplasm (figure 1b). When we screened our collection of mutants in this way we found other mutated forms mapping in the area of Lys-128 gave rise to both nuclear and cytoplasmic staining with individual cells displaying different proportions of staining in the two compartments (figure 1c).

In this way two sets of data could be obtained; the percentage of cells with nuclear fluorescence and the percentage with residual cytoplasmic fluorescence. Table 1 records the data obtained for many of the point mutants we have constructed in the region 121–137 together with some multiple mutants and some deletion mutants. Included in the table are some new mutants, particularly one (A1) in which Lys-128 has been converted to Arg, and a double mutant (PK1) in which we have attempted to reactivate the defective nuclear location signal of the d10 cytoplasmic mutant.

The importance of residue 128 has already been established in studies on d10 Large T (Kalderon et al. 1984a) and PARA-cT which has Asn at this position (Langford & Butel 1984). The data in table 1 emphasize the importance of residue 128 and also set limits on the nuclear location signal.

Residue 128

The importance of the amino acid at residue 128 is highlighted by the finding that all point mutations in the surrounding residues still give rise to nuclear fluorescence in 100% of the injected cells (table 1, column N). Thus, although nuclear location may be defective in some of the latter mutants, at least some Large T is still present in the nucleus. By contrast, alteration of Lys-128 to Thr, Ile, Met, Asn and Gln generates a Large T species that appears by immunofluorescence to be located almost exclusively in the cytoplasm. Only when Lys-128 is converted to Arg is a nuclear signal detected, and even then localization is defective with some cells displaying strong cytoplasmic fluorescence also observed. The implication of this result is that residue 128 is crucially important for nuclear localization of SV40 Large T, and a basic amino acid at this position is required with lysine much the preferred residue.

Surrounding region

Many single and multiple amino acid changes have been introduced into the basic tract including Lys-128 and the immediately surrounding region. Although in all cases nuclear Large T was detected, a variable amount of cytoplasmic Large T was also present in many cases. This phenotype of defective nuclear localization was caused only by mutations affecting residues between 125 and 132 (table 1, column C). Mutations outside this region had no discernible affect on subcellular location.
**Table 1. The Effect of Different Mutations within the Nuclear Location Signal on Subcellular Distribution of SV40 Large T**

(Figures above the amino acids indicate the residue number in wild-type Large T. N and C indicate the percentage of cells displaying nuclear (N) or cytoplasmic (C) Large T immunofluorescence.) The ability of virus constructed from the mutant plasmids to replicate in CV1 cells is recorded under the column heading R. +, Plaques of wild-type size were visible after 10 days; —, no plaques were detected after four weeks incubation. Plaque formation after a short delay (less than four days) relative to wild type is denoted by (D) and after a long delay (about 10 days) by D.

| Gly | His | Ser | Thr | Pro | Pro | Lys | Lys | Arg | Lys | Val | Glu | Asp | Pro | Lys | Asp | Plaque Formation |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----------------|
| 121 | 122 | 123 | 124 | 125 | 126 | 127 | 128 | 129 | 130 | 131 | 132 | 133 | 134 | 135 | 136 | 137 |
| Gly | Ser | Thr | Pro | Lys | Lys | Arg | Lys | Val | Glu | Asp | Pro | Lys | Asp | Wt. | C | R |
| 100 | 0   |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |

<table>
<thead>
<tr>
<th>N</th>
<th>C</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

A. E. Smith and others
The nuclear location signal

It is interesting that the substitution of any single amino acid by another leads to a variable effect on nuclear location depending on the nature of the substituent (table 1). Thus, when Lys-131 is converted to Thr (d27) 83 % of cells show cytoplasmic staining, when it is converted to Met (A8) 22 % of cells show cytoplasmic staining, whereas when it is converted to Arg (X8) only 9 % show this effect. A similar effect is seen when Lys-129 is converted to each of these amino acids. The proportion of cells exhibiting nuclear fluorescence varies if different time points are taken following microinjection as the defect in location appears to be kinetic. Nevertheless, these data may suggest that there is a hierarchy of favoured amino acids in the region around Lys-128. Perhaps this means that the region is recognized as some particular configuration or structure rather than simply as a basic tract.

The limits of the effective nuclear location signal in Large T can be defined by deletion and point mutations as the residues farthest away from Lys-128 for which alteration leads to some cytoplasmic fluorescence. The data from microinjection assays (table 1) indicate that Pro-125 defines the amino terminal end of the required structure (T23-L7, D29, D19, D30) and Val-132 the carboxy terminal end (W30, W38). These limits correspond almost exactly with those deduced in our earlier studies with a collection of cells transformed by mutants with slightly larger deletions (Kalderon et al. 1984a).

Biochemical properties of cytoplasmic Large T

The cytoplasmic Large T encoded by the d10 mutant is not defective in any known biochemical activities that have been tested in vitro (Paucha et al. 1985b). Thus, it binds to SV40 origin DNA and to DNA cellulose apparently normally. It also complexes with NV-T or p53, the transformation-associated cellular protein. The d10 virus does not replicate and the Large T it encodes is over-produced. The cause of this aberrant behaviour has not been proven. However, both these properties would be predicted if d10 Large T were unable to enter the nucleus to stimulate replication and transcription of viral DNA and to autoregulate its own synthesis. Thus, the biochemical properties of at least d10 Large T do not suggest the defect in nuclear location results from an inability to bind to a nuclear component. This argues against a ‘diffuse-and-bind’ model for the nuclear location of Large T.

Duplications and transposition of nuclear location signal in Large T

It seems unlikely that the cytoplasmic mutants have all acquired an affinity for a cytoplasmic component, since a number of different amino acids have been introduced at position 128 and the cytoplasmic phenotype is shared with several deletion mutants. We favoured the interpretation that the region around Lys-128 is part of a nuclear location signal that positively promotes entry to the nucleus, and that it is rendered defective by the d10 mutation (Kalderon et al. 1984a, b).

To test this hypothesis, Large T species were constructed containing two copies
of the amino acid sequences between residues 116 and 135 (Kalderon et al. 1984b). d10 and wild-type sequences were used and placed in all four possible combinations. The results obtained when the subcellular location of the chimeric Large T species was examined were unequivocal. Whenever a wild-type copy of the sequence between 116 and 135 was present the Large T was nuclear.

A more rigorous test was to attach the sequence around Lys-128 to the amino terminus of a Large T variant (S11–S33) which lacks amino acids 127–132 and which consequently encodes a cytoplasmic Large T. Plasmids were engineered so that either a wild-type or a d10 version of amino acids 127–147 was placed at the amino terminus of the mutant protein. Large T with a transposed wild-type sequence was predominantly nuclear, whereas the version containing the d10 sequence remained cytoplasmic.

The most rigorous test to demonstrate that the sequence from SV40 Large T strongly suggest that the sequence around Lys-128 does not interact with other parts of the Large T molecule but functions as an independent element that acts positively to promote nuclear location of Large T.

**Transposition of the nuclear location signal to other proteins**

The most rigorous test to demonstrate that the sequences from SV40 Large T around Lys-128 can act as an autonomous element was to translocate the sequence onto a protein that would otherwise be present in the cytoplasm and test whether it was able to relocate the protein to the nucleus. We reasoned that to do this would require that the sequence was attached at a position where it would be accessible, for example at the amino terminus. It also required a test protein with certain properties. It should be normally cytoplasmic or if its subcellular location is not known (as for a prokaryotic protein) it should be of sufficiently high molecular mass that it is unlikely to diffuse into the nucleus. Also cloned DNA encoding the protein and specific antibodies against the protein must be available. We found surprisingly few proteins that appeared to meet these requirements. Two that we have used are *Escherichia coli* β-galactosidase and chicken muscle type M1 pyruvate kinase (Kalderon et al. 1984b).

In both cases chimeric proteins were made that contained varying amounts of the SV40 Large T coding region attached to a fixed position near the amino terminus of the test protein. When 4 or 116 amino acids from Large T were attached at positions 6 and 17 of β-galactosidase and pyruvate kinase respectively, the distribution of proteins remained unaltered. However, addition of residues 1–136 resulted in the test proteins being located exclusively in the nucleus (figure 1e) and this effect was specific because when the equivalent d10 fragment encoding Thr at position 128 was used the subcellular distribution remained cytoplasmic (figure 1d).

Further experiments were designed to define the minimal sequences from SV40 Large T that could be attached at the amino terminus of pyruvate kinase and still result in the nuclear localization of the protein. To do this, small fragments of DNA encompassing the region around Lys-128 were cloned and successively shortened
The nuclear location signal

from adjacent restriction enzyme sites using Bal 31 nuclease. After addition of appropriate linkers these fragments were transposed to pyruvate kinase and the resulting plasmid DNA microinjected into Vero cells. Assuming the contribution made by the four amino acids from the amino terminus of SV40 Large T and those encoded by the linkers used in the construction is negligible, the results showed that the sequence Pro-Lys-Lys-Arg-Lys-Val can promote nuclear location pyruvate kinase (figure 1f).

It is difficult to refine the required sequences further because the contribution made by the linker coded amino acids cannot be ignored. Thus, the sequence Pro-Arg-Gly-Lys (128)-Lys-Arg-Lys-Val-(XR15) yields nuclear pyruvate kinase, whereas Ser-Ser-Arg-Arg (130)-Lys-Val-(XR24) does not. The sequence Ser-Ser-Arg-Lys (129)-Arg-Lys-Val-(XR12) gives rise to partly nuclear pyruvate kinase. In XR12 the position equivalent to Lys-128 is occupied by an Arg residue which we already know is the only amino acid able to replace Lys at this position (A1) without totally preventing nuclear location. This probably explains the partial ability of the XR12 truncated signal to function. Furthermore, our earlier experiments have already established the prime importance of residue 128. The positions equivalent to Pro-Pro-Lys at 125–127 may or may not be essential since the X15 construct which is the shortest that remains fully functional has a linker-coded proline at the position equivalent to 125 and the other positions (Pro-Lys) are replaced by Arg-Gly which are somewhat similar, in that they constitute a basic amino acid and a helix breaker, but in reverse order. The carboxy terminal end of the signal is more straightforward since Pro-Lys-Lys-Arg-Lys-Val is functional and Pro-Lys-Lys-Arg-Lys-Arg is not, indicating that the functional end-point lies at either position 131 or 132.

The foregoing discussion of course only applies to the requirements of the SV40 Large T nuclear location signal when positioned at a single location within pyruvate kinase. The requirements may vary when the signal is located at different sites within a given protein. Nevertheless, it is striking that the limits of the functional sequence in Large T and of the functional transposed sequence on pyruvate kinase are in very close agreement.

Computer search for signal in other proteins

The data presented above show unequivocally that the sequence Pro-Pro-Lys-Lys-Arg-Lys-Val can function as a nuclear location signal. Assuming that the signal is used as a general though not necessarily unique mechanism to gain access to the nucleus in eukaryotic cells, it follows that at least some other nuclear proteins should contain a similar signal. A preliminary search revealed that an identical sequence is found in only one other protein, that is, BK virus Large T, which in any case is very closely related to SV40 Large T. Thus, if there is a nuclear location signal in other proteins, it must be of different primary sequence to the SV40 Large T prototype. This is perhaps not unexpected in that we have already shown that many single amino acid changes can be introduced into the SV40 signal without drastic effect on function (Kalderon et al. 1984a, table 1). We therefore
conducted a computer search for related putative nuclear location signals in other proteins.

The search used was a modification (Greer et al. 1985) of the method described by Korn et al. (1985) and made use of the data base provided by the U.S. National Biomedical Research Foundation (released 20 January 1984) which contains 2511 protein sequences.

We first looked for sequences containing Pro or Pro-X followed by five basic amino acids. However, these revealed only six and twenty proteins, respectively, containing these sequences and almost half of these were histones or protamines. Since we knew of several other nuclear proteins in the data base that were candidates to contain a nuclear location signal, we progressively relaxed the requirements for the sequence both by reducing the homology required to only 75% to allow non-basic amino acids in the basic tract and by shortening the sequence of basic residues to four and three.

The searches for Pro or Pro-X followed by three basic residues did select a number of sequences from nuclear proteins including, for example, the sequence in polyoma virus Large T considered homologous to the SV40 Large T nuclear location signal (Tooze 1981). It also selected a number of other interesting candidates, for instance several virion proteins known to be nuclear (for example SV40 VP1 and VP2). However, several hundreds of other such sequences were detected and many were in non-nuclear proteins and in prokaryotic proteins and a very large number were present in histones.

We also searched for basic tracts preceded by Gly rather than Pro, since these amino acids both function as helix breakers in proteins and Hereford and colleagues (Hall et al. 1984; and personal communication) have suggested that in yeast Pro and Gly might both function in this position. This search selected further possible candidates. Finally, we screened a few proteins of particular interest (p53, adenovirus E1a, myc, cdc28, src, ras) seeking in the known nuclear proteins sequences that might be nuclear location signals, even though they might not fit the emerging consensus derived from the earlier searches.

Table 2 is a selection of sequences related to the SV40 Large T prototype. It is divided into two parts: (a) those that might act as nuclear location signals and (b) those that we feel are unlikely to act in this way.

(a) Possible candidates

The adenovirus 72 kDa DNA-binding protein, SV40 VP2/3, polyoma VP2/3, bovine papilloma virus (BPV) LI and EI, parvovirus VP2 and minute mouse virus (MMV) VP2 all contain sequences with good homology to the SV40 prototype. By using the alignment given in table 1 all contain Pro in position 2 or 3 relative to the Lys-128 equivalent and all contain at least four basic residues. It is striking that three of the sequences are immediately adjacent to the carboxy terminus of the proteins and two are present at the amino terminus. Presumably, basic sequences in these positions are very likely to be exposed in the overall structure of the proteins. Since all these proteins are known to be or are very likely to be nuclear proteins, we predict these sequences are nuclear location signals.

Polyoma virus VP1 has a possible nuclear location signal at its amino terminus
The nuclear location signal

but in this case it contains only three contiguous basic amino acids (Table 2). SV40 VP1 has a very similar sequence but it also contains a second possible signal, with three non-contiguous basic amino acids, a few residues downstream. The latter sequence is not conserved in polyoma virus VP1.

Polyoma virus Large T contains up to five possible nuclear location signals of varying fit to the SV40 Large T prototype. The sequences of SV40 and polyoma virus DNAs are considered homologous in the region that codes in SV40 for the nuclear location signal. However, the signal itself is only poorly conserved (Pro-Pro-Lys-Lys-Ala-Arg-Glu-Asp) in polyoma virus Large T.

It is possible that sequences more distantly related to the prototype can function as nuclear location signals in proteins such as polyoma virus Large T and SV40 VP1, but to do so they have to be very exposed in the overall structure of the protein. This might explain their presence at the amino terminus in polyoma virus and SV40 VP1 and would provide a mechanism to distinguish genuine, though short, signals from the many other copies of similar sequences present in non-nuclear proteins. An alternative explanation is that shorter basic tracts do function as nuclear location signals but only when present in multiple copies. These possibilities can be tested experimentally by placing different putative signals at different positions in cytoplasmic proteins and measuring their effect on subcellular location and by mutating the relevant sequences in different nuclear proteins.

When Gly is selected instead of Pro, a few further interesting candidates with four contiguous basic residues were selected in BPV E2, the SV40 agnogene product and in the Drosophila 70 kDa heat-shock protein. Since these are known or very likely to be nuclear proteins, these sequences, shown in table 2a, may also be nuclear location signals.

A putative nuclear location sequence is present in p53, the transformation associated protein which interacts with SV40 Large T. Chicken c-myc contains a tract of four Arg residues only 20 amino acids from its carboxy terminus, but the tract is not proceeded by Pro or Gly. Similarly, the adenovirus E1a protein has three basic amino acids within the C-terminal six residues but has no Pro or Gly preceding them. It is questionable whether the last two examples represent nuclear location signals. It is possible that c-myc and E1a enter the nucleus by a different mechanism, for example, by binding to another protein that carries them to the nucleus, or that they are present in another subcompartment of the nucleus, access to which requires a different signal.

Most of the putative nuclear location signals shown in table 2 are present in viral proteins. We believe this reflects the fact that the sequences of more viral nuclear proteins are present in the data-base than of their cellular counterparts. We expect that as the sequences of more nuclear proteins become available more putative signals in cellular proteins will emerge.

(b) Unlikely candidates

Table 2b lists a small selection of the sequences that fit, to varying degrees, the prototype nuclear location sequence, but which we do not believe function as such for various reasons.
TABLE 2. SELECTED AMINO ACID SEQUENCES WITH HOMOLOGY TO THE SV40 LARGE T PROTOTYPE NUCLEAR LOCATION SIGNAL

(Details of all the sequences are in the U.S. National Biomedical Research Foundation database. Columns on the right indicate whether the protein is nuclear, whether it interacts with nucleic acid, whether the sequence given is from a more extensive basic region and whether we consider the sequence a likely candidate nuclear location signal. Symbols in parentheses indicate a likely property but which is not yet established, and ? indicates an unknown property.)

<table>
<thead>
<tr>
<th>Protein</th>
<th>Sequence</th>
<th>Interacts</th>
<th>Nucleic Acid</th>
<th>Basic Region</th>
<th>Candidate</th>
</tr>
</thead>
<tbody>
<tr>
<td>SV40</td>
<td>Large T</td>
<td>P.P-K-K-K-R-K-V</td>
<td>+</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>BK</td>
<td>Large T</td>
<td>P.P-K-K-K-R-K-V</td>
<td>+</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>adenovirus</td>
<td>72K DBP</td>
<td>A.P-K-K-K-K-R</td>
<td>+</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>SV40</td>
<td>VP2/3</td>
<td>P.N-K-K-K-R-K-L</td>
<td>+</td>
<td>(+)</td>
<td>-</td>
</tr>
<tr>
<td>polyoma</td>
<td>VP2/3</td>
<td>P.Q-K-K-K-R-R-L&lt;sub&gt;COOH&lt;/sub&gt;</td>
<td>+</td>
<td>(+)</td>
<td>-</td>
</tr>
<tr>
<td>BPV</td>
<td>L1</td>
<td>P.A-K-K-K-K-K&lt;sub&gt;COOH&lt;/sub&gt;</td>
<td>(+)</td>
<td>(+)</td>
<td>-</td>
</tr>
<tr>
<td>BPV</td>
<td>E1</td>
<td>T.P.V-K-R-R-K-S</td>
<td>(+)</td>
<td>?</td>
<td>-</td>
</tr>
<tr>
<td>HPV</td>
<td>VP2</td>
<td>P.S-L-R-K-R-R-A-Y&lt;sub&gt;COOH&lt;/sub&gt;</td>
<td>(+)</td>
<td>(+)</td>
<td>-</td>
</tr>
<tr>
<td>parvovirus</td>
<td>VP2</td>
<td>P.P-A-K-R-K-R-G</td>
<td>(+)</td>
<td>(+)</td>
<td>-</td>
</tr>
<tr>
<td>MMV</td>
<td>VP2</td>
<td>P.P-A-K-R-A-K-R</td>
<td>+</td>
<td>(+)</td>
<td>-</td>
</tr>
<tr>
<td>polyoma</td>
<td>VP1</td>
<td>M-A-P-K-R-K-S-G</td>
<td>+</td>
<td>(+)</td>
<td>-</td>
</tr>
<tr>
<td>SV40</td>
<td>VP1</td>
<td>A-P-T-K-R-K-G-S</td>
<td>+</td>
<td>(+)</td>
<td>-</td>
</tr>
<tr>
<td>polyoma</td>
<td>Large T</td>
<td>P.P-K-A-R-E-D</td>
<td>+</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A-G-R-R-R-N-P-C</td>
<td>.</td>
<td>.</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>P-V-S-R-K-R-P</td>
<td>.</td>
<td>.</td>
<td>-</td>
</tr>
<tr>
<td>BPV</td>
<td>E2</td>
<td>Q.G-R-K-K-R-S-S-R</td>
<td>(+)</td>
<td>?</td>
<td>-</td>
</tr>
<tr>
<td>SV40</td>
<td>Agno</td>
<td>V.D.G-K-R-K-K-P</td>
<td>+</td>
<td>+</td>
<td>-</td>
</tr>
</tbody>
</table>
### TABLE 2 (cont.)

<table>
<thead>
<tr>
<th>Protein</th>
<th>Sequence</th>
<th>Nuclear acid</th>
<th>Basic region</th>
<th>Candidate</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) Possible candidates</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Drosophila 70K HSP</td>
<td>K-A-G-K-R-K-K-P</td>
<td>+</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>mouse p53</td>
<td>P-P-Q-K-K-K-P-L</td>
<td>(+)</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>chicken c-myc</td>
<td>K-E-Q-L-R-R-R-R-E</td>
<td>+</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>adenovirus E1A</td>
<td>L-S-V-K-R-P-R-C-N_COOH</td>
<td>+</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>(b) Unlikely candidates</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SBMV CP</td>
<td>P-P-R-R-K-R-R</td>
<td>(-)</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>SFV CP</td>
<td>P-K-P-K-K-K-K</td>
<td>(-)</td>
<td>(+)</td>
<td>+</td>
</tr>
<tr>
<td>Polio type I VP3</td>
<td>P-P-K-K-R-K-E</td>
<td>(-)</td>
<td>(+)</td>
<td>-</td>
</tr>
<tr>
<td>type II VP3</td>
<td>P-P-K-S-R-K-E</td>
<td>(-)</td>
<td>(+)</td>
<td>-</td>
</tr>
<tr>
<td>type III VP3</td>
<td>P-P-T-S-R-K-E</td>
<td>(-)</td>
<td>(+)</td>
<td>-</td>
</tr>
<tr>
<td>Hepatitis CAg</td>
<td>S-P-R-R-R-R-S</td>
<td>(+)</td>
<td>(+)</td>
<td>+</td>
</tr>
<tr>
<td>Tuna Protamine</td>
<td>P-R-R-R-R-R</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Sea Urchin H1</td>
<td>S-P-R-K-A-K</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
</tbody>
</table>

One group of proteins with extensive basic tracts preceded by Pro are capsid proteins from viruses that replicate predominantly, if not exclusively, in the cytoplasm of their hosts. There is little reason to expect these proteins to have nuclear location signals. Sometimes (for example, cauliflower mosaic virus (CMV) core protein (CP), hepatitis core antigen (CAg)) the tracts are part of even longer basic regions similar to those found in histones or protamines (table 2b). All such
sequences may be involved in interactions with nucleic acids and function in packaging of the virions. Indeed, the crystal structure of southern bean mosaic virus (SBMV) predicts that the amino terminal 50 residues which includes the homologous region does indeed interact with viral nucleic acid (Abad-Zapatero et al. 1980). Thus sequences in this category, although related to the nuclear location signal, may be inaccessible and consequently unable to function as such. Further evidence against an important role for the sequence in poliovirus, which incidentally can replicate in enucleate cells, is that the basic tract in VP3 is not conserved in different virus types even though the surrounding residues are (table 2b).

Another group of proteins that contain homology with the nuclear location signal are protamines and histones. Although these are nuclear proteins, they may be small enough to enter the nucleus by diffusion and concentrate there because of their high affinity for DNA. In some cases the sequences are present at the amino terminus (tuna protamine) and sometimes at the carboxy terminus (sea urchin histone H1). In other cases the sequences are repeated many times. For example, Pro followed by at least four basic residues occurs six times in sea urchin histone H1. However, the occurrence of such repeats is almost inevitable given the composition of these proteins. At present it is difficult to judge whether these sequences act as nuclear location signals and, because of their common occurrence, it will be difficult to establish this experimentally. Presently we tend to discount them.

Another feature of the sequences we have rejected as putative nuclear location signals and most of which are not shown in table 2b, is a predominance of Arg rather than Lys. In the signals that we feel most confident about (table 2a) Lys is much more common than Arg.

We believe that there will be many proteins that contain sequences that fit whatever consensus sequence for the nuclear location signal finally emerges, but which nevertheless do not function as such. It would follow that although such a sequence in a protein may be necessary and sufficient for entry to the nucleus by one pathway, the presence of such a sequence does not mean that the protein will necessarily be nuclear. Presumably, such sequences could be unavailable or inaccessible either because they interact with some other part of their protein structure or because they interact with high affinity with some other component such as virus capsid proteins or nucleic acid. In other cases another signal, for example for membrane insertion, could take precedence by acting in the nascent protein. For these reasons we predict that the nuclear location sequence will function only in some circumstances in much the same way as the prototype sequence Asn X. Thr acts as a glycosylation signal only at some positions in some proteins. A consequence of this is that it will probably not be possible to distinguish nuclear and cytoplasmic proteins on the basis of amino acid sequence alone.

**Nuclear location domains**

One of the most surprising features of the nuclear location signal defined here is its small size, around seven amino acids. What might be termed the nuclear location domain of proteins must either be very small relative to the overall
The nuclear location signal

SV40

\[
\begin{array}{cccccccccccc}
315 & 316 & 317 & 318 & 319 & 320 & 321 & 322 & 323 & 324 & 325 \\
Asp & Gly & Pro & Asn & Lys & Lys & Lys & Arg & Lys & Leu & Ser \\
GATGGCCCCAACAAAAAGAAAAGGAAGTTGTCC & & & & & & & & & & \\
\end{array}
\]

VP1

\[
\begin{array}{cccccccccccc}
Met & Ala & Pro & Thr & Arg & Lys & Gly & Ser & Cys & & \\
1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & \\
\end{array}
\]

polyoma virus

\[
\begin{array}{cccccccccccc}
310 & 311 & 314 & 315 & 316 & 317 & 318 & 319 \\
Asp & Gly & Pro & Gln & Lys & Lys & Arg & Arg Lys Arg Leu COOH \\
GATGGCCCCCAAAAGAAAAAGCGGCGTCTCA & & & & & & & & & & \\
\end{array}
\]

VP1

\[
\begin{array}{cccccccccccc}
Met & Ala & Pro & Lys & Arg & Lys & Ser & Gly & Val & Ser & \\
1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & \\
\end{array}
\]

Figure 2. Position of putative nuclear location signal in papovavirus capsid proteins. Solid bars indicated the putative nuclear location signals.

structure of the average protein or it must be part of another larger structural domain.

Figure 2 shows the arrangements of DNA sequences encoding the putative SV40 and polyoma virus VP2 nuclear location domains. These are near or at the carboxy terminus of their respective proteins. The same DNA sequences encode, in a different reading frame, the amino terminus of VP1 and includes a potential nuclear location signal for this protein. We have argued above that it is likely that nuclear location signals will be present at the termini of proteins, because in this position they should be exposed and place no constraints on the folding of the remainder of molecule. The arrangement shown in figure 2 is a particularly striking example.

Our mutagenic studies of the region of SV40 Large T between residues 106 and 158 and those of others (summarized in Paucha et al. 1985a) allow a detailed functional domain structure for this region to be constructed. Figure 3 shows that at least three domains appear to be adjacent to one another in this region; a highly phosphorylated region (ca. 105–125) (Scheitmann et al. 1982); a DNA binding domain (139–ca. 220) (reviewed in Paucha et al. 1985a) and between them the nuclear location domain (125–132). It is likely that the nuclear location signal is very exposed in Large T because the bond between Arg-130 and Lys-131 is one of the most sensitive in the molecule to limiting amounts of protease (Schwyzer et al. 1980). Whether this forms an exposed ‘hinge’ in the molecule or is part of a larger exposed structure that is accessible to kinases, proteases and to DNA remains unknown.
FIGURE 3. Position of nuclear location signal in SV40 Large T. Boxes indicate the location of the putative domains on SV40 Large T. Crosses within indicate the location of a point mutant that abolishes or reduces the activity of the particular domain.

MECHANISMS OF NUCLEAR LOCATION

Throughout this review we have avoided describing the process that results in Large T being transferred to the nucleus as 'nuclear transport', because this implies we know the mechanism involved. This is not the case.

Our current hypothesis is that the signal identified here is recognized directly by a protein present in the cytoplasm, the nuclear membrane or nuclear pores and this acts as a carrier to transport the protein in question into the nucleus. However, we are not aware of any evidence in favour of this mechanism. The availability of the mutants described here should make possible the development of cell-free systems capable of such transport, and provide good controls with which to test the specificity of such systems.

Another major issue that requires further study is the generality of the mechanism which makes use of the signal defined here. We feel confident that it will be used by a number of large nuclear proteins but whether it applies to other smaller proteins with high affinity for nuclear components remains unclear.

NUCLEAR LOCATION AND TRANSFORMATION

The purpose of the mutagenic study that produced the initial nuclear location-defective mutants was to identify functions essential for transformation. The unexpected phenotype of the d10 and related mutants allows us to probe the requirement for nuclear location for the transformation of established cell lines and the immortalization of primary cells.

Our initial studies show that the d10 mutant retains the ability to transform Rat-1 cells as judged by focus formation in high serum (Kalderon et al. 1984a). Such d10 transformed cells also have the ability to grow in semi-solid medium
suggesting that they are more than minimally transformed. The implication of this result is that transformation of Rat-1 cells by SV40 may be a non-nuclear event catalysed by the membrane associated forms of Large T characterized by Klockmann & Deppert (1983).

However, although d10 transformed cells appear to have very little or no nuclear Large T as judged by immunofluorescence (Kalderon et al. 1984a) and by the reduced ability of the d10 Large T to autoregulate its synthesis (Paucha et al. 1985b), further studies of subcellular and subnuclear location are necessary. More detailed studies of the biology of d10 transformed Rat-1 cells are also required. If the current view is correct that transformation of primary cells is a multi-step process, first involving immortalization which occurs as a nuclear event (Land et al. 1983), we might further predict that d10 Large T would be unable to transform primary cells. The availability of the various mutants described here makes many of these predictions very easy to test experimentally.

References


Bonner, W. M. 1975 Protein migration into nuclei. II. Frog oocyte nuclei accumulate a class of microinjected oocyte nuclear proteins and exclude a class of microinjected oocyte cytoplasmic proteins. J. Cell Biol. 64, 431–437.


Paucha, E., Kalderon, D., Harvey, R. W. & Smith, A. E. 1985a The SV40 origin DNA binding domain on large-T antigen. (Submitted.)

Paucha, E., Kalderon, D., Richardson, W. D., Harvey, R. W. & Smith, A. E. 1985b The abnormal location of cytoplasmic SV40 large-T is not caused by failure to bind to DNA or to p53. (Submitted.)


Figure 1. Subcellular location of SV40 large T and pyruvate kinase variants. Subconfluent Vero (African green monkey) cells were microinjected with plasmid DNA encoding wild-type or mutant SV40 Large T (a)-(c), or fusion proteins containing portions of Large T linked to amino acids 17-529 of pyruvate kinase (PK) (d)-(f). Approximately 18 h later the cells were fixed, and proteins made visible by immunofluorescence microscopy. SV40 Large T was detected by using as first antibody a monoclonal directed against its carboxy-terminal region. (a) Wild-type Large T; (b) d10 mutant Large T (Lys 128→Thr); (c) d1 mutant Large T (Lys 129→Met, Lys 131→Thr). Large T–PK fusion proteins were labelled with a polyclonal antibody raised in rabbits against purified PK. (d) Amino acids 1-136 of d10 mutant Large T (Lys 128→Thr) fused to PK; (e) amino acids 1-136 of wild-type Large T