INTRODUCTION

One of the most conspicuous features of the eukaryotic cell is its compartmentalization. Numerous organelles and domains subdivide the anatomy of the cell. Moreover, most of these can be further subdivided into well differentiated regions or structures. The physiological relevance of such compartmentalisation is paramount. Every major cellular activity can be assigned to one or more subcellular compartments. Furthermore, the intricate regulatory networks which operate within eukaryotic cells greatly rely on the differential compartmentalisation of their components. The close relationship between subcellular localisation and function is such that determining the preferential localisation of a protein is often an essential step towards determining its function. Typically, the identification of proteins that are localised in a given compartment requires a very lengthy procedure that relies on the preparative isolation of such a compartment through cell fractionation techniques. This approach presents several limitations. Firstly, it is totally dependent upon the yield and enrichment that can be achieved by these cell fractionation techniques. Secondly, cell fractionation can lead to the loss of peripheral, yet functionally relevant components. Finally, and more importantly, it can only be applied to those compartments that are amenable to cell fractionation. The relative amount and the biochemical characteristics of every single protein can impose an additional burden.

The genomic revolution is providing an ever increasing number of sequences with unknown functions. To match this wealth of sequence data, high throughput methods to generate functional information, of which subcellular localisation is an essential part, are needed. An attempt to this end was provided by the method developed to identify nuclear markers in fission yeast (Sawin and Nurse, 1996). However, with the sequencing of higher eukaryotes, a similar technique is needed for high efficiency determination of subcellular localisation in these organisms. To identify protein motifs that are able to target proteins to a distinct localisation in cultured cell lines derived from higher eukaryotes we have developed a rapid procedure termed Motif Trap (MT). The method is based on libraries made of DNA fragments fused to GFP. Following transfection with a MT library, the cells that display a GFP localisation of interest are isolated and the corresponding cDNA fragment that encodes the targeting motif cloned by RT-PCR. These can then be used to identify the corresponding full-length cDNAs. Using MT, we have identified patterns of GFP localisation which correspond to every major organelle and compartment. We have shown that MT is useful to identify new sequences that determine subcellular localisation as well as known targeting motifs.

SUMMARY

We have developed a rapid procedure termed Motif Trap (MT) to identify protein motifs that are able to target proteins to a distinct subcellular localisation in eukaryotic cells. By expressing random DNA fragments fused to green fluorescent protein (GFP), individual cells with the GFP localisation of interest are readily isolated allowing for the expressed DNA fragments to be cloned by RT-PCR. These can then be used to identify the corresponding full-length cDNAs. Using MT, we have identified patterns of GFP localisation which correspond to every major organelle and compartment. We have shown that MT is useful to identify new sequences that determine subcellular localisation as well as known targeting motifs.

MATERIALS AND METHODS

Construction of the MT vector (MT#1)

Using primers A (CATGTTGGCGGCGGTACCGTGAG) and B (GCCCGGGCGTGAGCAAGGGCGAG) we modified pEGFP-N1 (Clontech) by PCR to introduce an SrfI site between nucleotides three and four of the GFP coding sequence. This insertion shifts the initial ATG codon of the GFP out of frame with the rest of the coding sequence. This ensures that only insert-containing plasmids will express GFP. PCR was carried out with the Expand High Fidelity PCR System (Boehringer). Oligo A also introduced a NotI site 10 nucleotides upstream of the GFP CDS. This insertion shifts the initial ATG codon of the GFP out of frame with the rest of the coding sequence. This ensures that only insert-containing plasmids will express GFP. PCR was carried out with the Expand High Fidelity PCR System (Boehringer). Oligo A also introduced a NotI site 10 nucleotides upstream of the GFP CDS. The PCR product was purified using the PCR Quiagquick Purification System (Qiagen), ligated with Rapid Ligation Kit (Boehringer) and used to transform Epicurian Coli XL-1-Blue (Stratagene), by heat-shock. Transformed cells were plated out in LB-Agar supplemented with 30 μg/ml kanamycin and incubated for 16 hours at 37°C. The modified vector was then isolated by minipreps and the NotI fragment subcloned into a pQE31 vector.
previously modified to introduce a NotI site between the BamH1 and KpnI sites with an adaptor made with oligos Not1-1b (GATC CGCGCCGCGTAC) and Not1-8 (GCGGCCGC). The resulting colonies were checked under a transilluminator to test the expression of GFP and the NotI fragment was then isolated from one of the colonies and subcloned into pEGFP-N1-Not, a modified version of pEGFP-N1 that carries an additional NotI site inserted in position 635-642.

Construction of the MT libraries
To construct the MmcDNA-MT library, cDNA was obtained from NIH/3T3 cells by random priming, purified with QIAEX II Gel Extraction Kit, cloned into the SrfI site of the MT#1 vector using the Rapid Ligation Kit (Boehringer) and transformed into E. coli XL1-Blue MR (Stratagene). Plating out a small aliquot of these cells, we estimated that the library contained about 420,000 clones of which 1.6% had no insert. The complete library was then plated out onto a sterile Nylon filter laid on a 24x24 cm plate containing LB supplemented with kanamycin, incubated at 30°C for 24 hours, replicated into another filter, and reincubated for 4 hours at 37°C. The DNA from the library was then purified with Plasmid Maxi kit (QIAGEN). The DmgDNA-MT library was prepared in the same way except that the inserts in this case were Drosophila genomic DNA fragments between 100 and 150 nt in length.

Transfection of HEK293 cells with the MT libraries and cloning of cells displaying localised GFP
HEK293 cells were transfected with the MmcDNA-MT library, GFP is shown in green. These cells were counterstained for DNA using propidium iodide (red). (B) Mitochondria; (C) centrosomal region (arrow); (D) cytokinesis furrow (arrow); (E) the chromosomes (arrows); (F) the mitotic spindle. We have not determined yet the subcellular localisation of GFP in the cells shown in G,H,I. Bars, 15 μm.
Identification of protein targeting motifs to sixteen hours after transfection to check for localised GFP using an inverted LEICA DMI-RBE microscope with a long distance 63× Fluorite objective. The position of the cells of interest was labelled with a diamond pen, and then cloned by a combination of manual cloning and serial dilutions, as described by Harlow and Lane (1988). In some cases, the cells were first cloned using a fluorescence-activated cell sorter (FACS) and the resulting clones were later analysed to determine the presence of localised GFP.

RESULTS AND DISCUSSION

The construction of a Motif Trap library and its use to identify protein motifs that can target proteins to different cellular compartments is summarised in Fig. 1. The source DNA, either genomic or cDNA, is fragmented and cloned into the MT vector. Upon transfection with a MT library, the cells that have taken up a plasmid carrying an insert that encodes a targeting signal can readily be identified by the localisation of GFP. These cells are then isolated by a combination of manual and serial dilution cloning and the corresponding cDNA fragments cloned by RT-PCR.

Between eight and ten hours after transfection of HEK293 cells with the MmcDNA-MT library, some cells start to express GFP and the first localisation patterns are recognisable (Fig. 2A). The percentage of cells transfected with a MT library that are expected to express GFP depends exclusively on the proportion of clones that carry a stop codon in frame with the GFP. This, in turn, depends on the size and base composition of the insert. Thus, the proportion of positive clones is expected to be high in libraries made with short inserts and should be minimum in MT libraries made of full-length cDNAs. In a typical transfection experiment with the MmcDNA-MT library, about 50% of the cells express GFP of which 20% display a distinct localisation of this reporter. Fig. 2B to I shows some of the GFP localisation patterns that we observed. Fig. 2B shows GFP specifically localised in the...
mitochondria as confirmed by counterstaining with the mitochondria-specific marker mitotracker (not shown). In the cell shown in Fig. 2C, GFP displays a fairly uniform distribution in the cytoplasm, but is significantly concentrated in a small area near the nucleus that corresponds to the centrosome (arrow) as revealed by counterstaining with a human autoimmune anti-centrosome antibody (not shown). Fig. 2D,E and F show mitotic cells from different GFP expressing lines. GFP can be seen to localise in the cytokinesis furrow (arrow; Fig. 2D), the chromosomes (Fig. 2E) and the mitotic spindle (Fig. 2F). GFP does not appear to be localised during interphase in these two cell lines. We have not yet determined the precise subcellular localisation of GFP in the cells shown in Fig. 2G,H and I. These are a few examples of the patterns of GFP localisation that we have observed. Using MT we have been able to identify cells with GFP localised in every major organelle and compartment.

To demonstrate the use of MT to isolate protein targeting motifs we have cloned and sequenced the DNA inserts from some of these cells. As expected, we have found sequences that correspond to known proteins and contain targeting signals which are consistent with the observed localisation of the GFP fusion. One of these is clone 02/11#22, (Fig. 3A,B). The GFP fusion in this cell line shows a distinct nucleolar localisation with a weak nuclear background. The insert from this line is identical to the fragment that spans between amino acids 62 and 131 of the mouse homologue of the HTLV-I tax responsive element binding protein TAXREB107 (Nacken et al., 1995). This fragment, contains a well defined bipartite nuclear localisation signal (KKKYSAAKTKVEKKKKKE) and meets the consensus of a nucleolus localisation signal. We have also found inserts that are new sequences which do not have a match in the databases. This is the case of clone 09/07#18 (Fig. 3C,D). These cells contain GFP that is tightly localised to the endoplasmic reticulum (ER), as shown by counterstaining with an antibody against the ER marker α-calnexin (not shown; Cannon and Helenius, 1999). The insert from this cell line encodes a peptide, 35 amino acids long and contains a predicted trans-membrane motif (PMSIFIQLY-FLFLFLGVIC) that may account for the ER localisation shown by the fusion protein (von Hejne, 1992). Indeed, this does not mean to say that the identified motif contains an active retention signal since the ER accumulation of the fusion protein could result from its failure to progress further into the secretory pathway. These observations demonstrate the use of MT to identify and clone protein fragments bearing specific targeting signals.

The main feature of MT is that it is not biased by the particular features of a given protein. Polypeptides containing a localisation signal are identified solely on the basis of their capability to drive GFP to a particular subcellular compartment. The abundance, physiological role, and possible regulation of the full-length protein is totally irrelevant. Moreover, MT is not limited to compartments that are amenable to cell fractionation. In principle, proteins belonging to any compartment that can be visualised by fluorescence microscopy could be identified with this method.

Nevertheless, several limitations apply to the Motif Trap approach. Since the MT vector generates N-terminal fusions with the reporter, it cannot be used to identify localisation signals that reside in the C-terminal end of the molecule such as CAAX motifs (Hancock et al., 1990) or proteins whose compartmentalisation relays on the subcellular localisation of their transcripts, which usually depends on the 3′UTR region (reviewed by Kislauskis and Singer, 1992). These limitations could be circumvented by a different MT vector designed to produce C-terminal fusions, but in this case the reporter gene would be expressed regardless of the presence of an insert and whether or not the insert carries an open reading frame. Consequently, a great many more cells would have to be scored to find clones that display a particular localisation of the reporter.

An additional limitation of the MT approach is due to the very nature of the subcellular localisation motifs. Given the short size and low complexity of many of these, stretches of amino acids that match a consensus, but do not act as targeting sequences are not uncommon. For instance, it has been estimated that a significant fraction of the proteins that carry a nuclear localisation signal are never found in the nucleus (Dingwall and Laskey 1991). Mitochondrial signal sequences (or pre-sequences) also fall into this category. Mitochondrial presequences have been defined as N-terminal peptides that are generally basic, 15-30 amino acids long, rich in hydroxilated amino acids, deficient in acidic amino acids, and can form an amphipilic α-helix or β-sheet (Roise and Schatz, 1988). To get an estimate of the incidence of cryptic mitochondrial signals and how could they affect the identification of mitochondrial proteins using MT, we made a library with small inserts of Drosophila genomic DNA that was then tested in human HEK293 cells. We used genomic DNA from Drosophila as a source or high complexity DNA which, when cut in small inserts, would render a nearly random collection of sequences. The proportion of clones from this library that displayed a mitochondrial localisation was very high, around 15% of the GFP expressing cells. Cloning and sequencing of seven of these inserts showed that none of them corresponds to known mitochondrial proteins. In fact, most of them corresponded to very short ORFs present in non-coding sequences like satellite DNA and rDNA (Fig. 4). Interestingly, despite the nearly random origin of these inserts, all of them fit very precisely the defined consensus for a mitochondrial pre-sequence: short, with a net positive charge, containing some hydroxilated amino acids and predicted to fold as an amphipilic α-helix, thus illustrating the power of the MT approach to define consensus localisation signals using nearly-random DNA.

Short coding sequences, just like non-coding sequences, are therefore expected to contain numerous cryptic subcellular localisation motifs that are normally silent due to different reasons. Probably, the most common one is their position within the folded protein which renders them inaccessible to the other components of the subcellular localisation machinery (Roise and Schatz, 1988). In addition, some potential localisation signals may fail to target because of the presence of other targeting signals that act in a dominant manner and recruit the protein to a different compartment (Dingwall and Laskey 1991). These cryptic targeting signals can be revealed by taking them out of their natural context in a fusion with a reporter protein. Moreover, short DNA fragments can also contain non-coding ORFs that may encode artefactual targeting sequences. Although non-naturally occurring sequences capable of targeting may have very interesting applications,
they can produce a bothersome background when naturally occurring targeted proteins are to be identified. The incidence of all these artefactual results can be considerably reduced by the use of MT libraries made with large CDNA inserts so that nearly full-length proteins are fused to the reporter. These protein fusions have a higher chance of folding, thus inactivating cryptic targeting signals. They also have more chances of carrying the more dominant targeting motif, the one that determines the normal localisation of the protein. Moreover, longer cDNAs will increase the frequency of stop codons in any of the incorrect ORFs thus decreasing their artefactual expression.

Finally, the MT approach is constrained by the limitations that apply to artificial protein expression in eukaryotic cells. The main source of artefactual results is overexpression. Overexpression can lead to the ectopic accumulation of a protein in one of the compartments it normally goes through to reach its final destination, to the production of aggregates, and, when overexpression results in protein levels which exceed the capability of the protein folding machinery, to the recruitment of the unfolded protein in the aggresome (Johnston et al., 1988).

Despite these potential limitations, that can be reasonably controlled by choosing the right vector, insert size and expression conditions, MT provides a rapid approach to identify localised proteins. There are several potential applications of MT that we can envisage will be developed in the future. MT may offer a powerful tool in the field of functional genomics as a high throughput method to determine the subcellular localisation of the fast growing number of sequences which are being generated by ongoing genome projects. Another application of MT is the identification of proteins that are differentially sorted in differentiating cells like epithelial cells that are induced to polarise or primary cultures of differentiating neurons (Dotti and Simons, 1990). Similarly, MT could be used to detect proteins which are specifically relocated under different stress conditions, following heat-shock for instance, or after infection with pathogens which result in a dramatic rearrangement of the architecture of the cell (Cudmore et al., 1997). Another interesting application of MT will be the cloning of interacting partners of a given protein by transfecting cells which contain the protein labelled with a fluorochrome that produces FRET (Cubitt et al., 1995) with the library’s reporter. Finally, systematic screenings with a MT library could be used to identify new domains within known organelles and compartments. An additional spin off of MT is the establishment of a large collection of cell lines that carry GFP at different locations.

While this paper was being reviewed, a related manuscript was published describing a visual screen of a GFP-fusion library (Rolls et al., 1999). Unlike the method described here, this screening was carried out by examining the patterns of GFP localisation in cells transfected with pools of clones generated by subdividing the library. Further subdivision into smaller pools and rescreening was used to isolate the clones responsible for a particular pattern of interest. Another difference with the MT method is that in the library used by Roll et al. (1999) the cDNAs were inserted after the 3’ end of the GFP coding sequence. Technical differences aside, the visual screen described by these authors and the MT approach are conceptually identical and represent the first two of what may soon become a large number of variations on the same theme.

We are indebted to M. Way, C. Dotti, S. Cohen, E. Karsenti, T. Nilsson, T. Hyman, R. Pepperkok and G. Griffiths for their encouragement and support as well for their critical reading of this manuscript. Some of the initial experiments that lead to the development of MT were carried out by L.A.B. in the laboratory of J. Jimenez. G. Alonso, suggested the use of the HEK293 cells. L.A.B. is recipient of a postdoctoral fellowship from the Spanish ‘Ministerio de Educacion’. Work in our laboratory is partially financed by EU grant CT960005.

REFERENCES


