Office de la Propriété Intellectuelle du Canada
Un organisme d'Industrie Canada

Canadian Intellectual Property Office
An agency of Industry Canada

CA 2367723 A1 2000/09/28

(21) 2 367 723

(12) DEMANDE DE BREVET CANADIEN
CANADIAN PATENT APPLICATION

(13) A1

(86) Date de dépôt PCT/PCT Filing Date: 2000/03/13
(87) Date publication PCT/PCT Publication Date: 2000/09/28
(85) Entrée phase nationale/National Entry: 2001/09/19
(86) N° demande PCT/PCT Application No.: US 2000/006560
(87) N° publication PCT/PCT Publication No.: 2000/056878
(30) Priorité/Priority: 1999/03/19 (60/125,126) US

(51) Cl.Int.^{7}/Int.Cl.^{7} C12N 15/10, C12N 9/90, C12N 15/85, C12N 15/70, C12Q 1/68, C12N 15/62, C07K 14/47

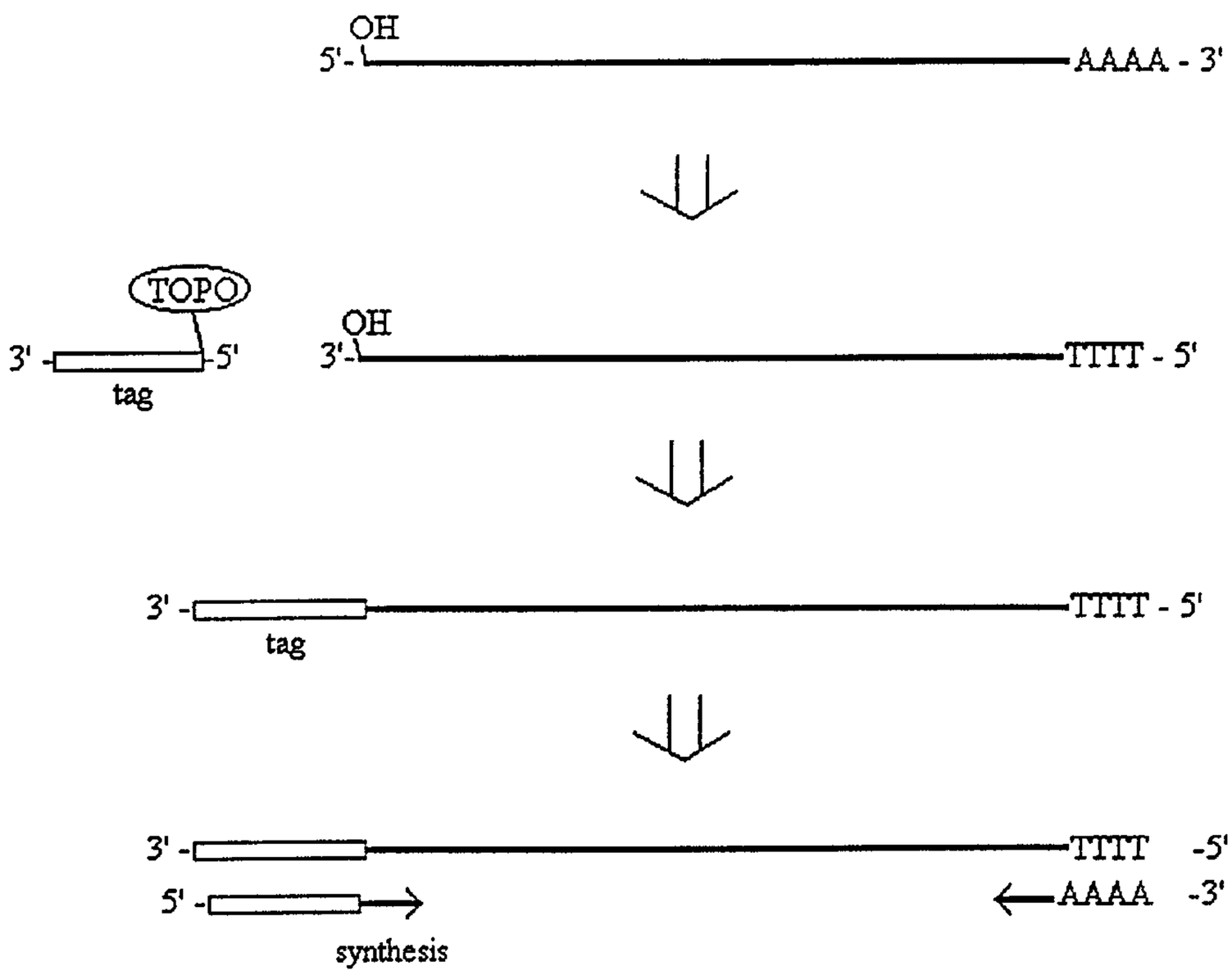
(71) Demandeur/Applicant:
INVITROGEN CORPORATION, US

(72) Inventeurs/Inventors:
PHELAN, DOROTHY, US;
MARCIL, ROBERT, US;
COMISKEY, JOHN D., US;
HEYMAN, JOHN A., US

(74) Agent: MBM & CO.

(54) Titre : PROCEDES D'OBTENTION DE SEQUENCES D'ACIDE NUCLEIQUE PLEINE LONGUEUR A L'AIDE DE E. COLI TOPOISOMERASE III ET SES HOMOLOGUES

(54) Title: METHODS OF OBTAINING FULL-LENGTH NUCLEIC ACID SEQUENCES USING E. COLI TOPOISOMERASE III AND ITS HOMOLOGS



(57) Abrégé/Abstract:
The invention disclosed herein comprises a method of obtaining full-length coding cDNA. The invention method comprises using isolated full-length mRNA to synthesize first strand cDNA, attaching a non-native tag to the cDNA and using the tagged first strand to synthesize second strand cDNA. The final cDNA can additionally be amplified and inserted into an expression vector. Also disclosed are isolated full-length coding cDNAs prepared using the method of the invention, as well as nucleic acid constructs containing the full-length coding cDNAs.

PCT

WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification [7] :
C12N 15/10, 15/62, 15/70, 15/85, 9/90, C07K 14/47, C12Q 1/68

A1

(11) International Publication Number: **WO 00/56878**

(43) International Publication Date: 28 September 2000 (28.09.00)

(21) International Application Number: PCT/US00/06560

(22) International Filing Date: 13 March 2000 (13.03.00)

(30) Priority Data:
60/125,126 19 March 1999 (19.03.99) US

(63) Related by Continuation (CON) or Continuation-in-Part (CIP) to Earlier Application
US 60/125,126 (CON)
Filed on 19 March 1999 (19.03.99)

(71) Applicant *(for all designated States except US)*: INVITROGEN CORPORATION [US/US]; 1600 Faraday Avenue, Carlsbad, CA 92008 (US).

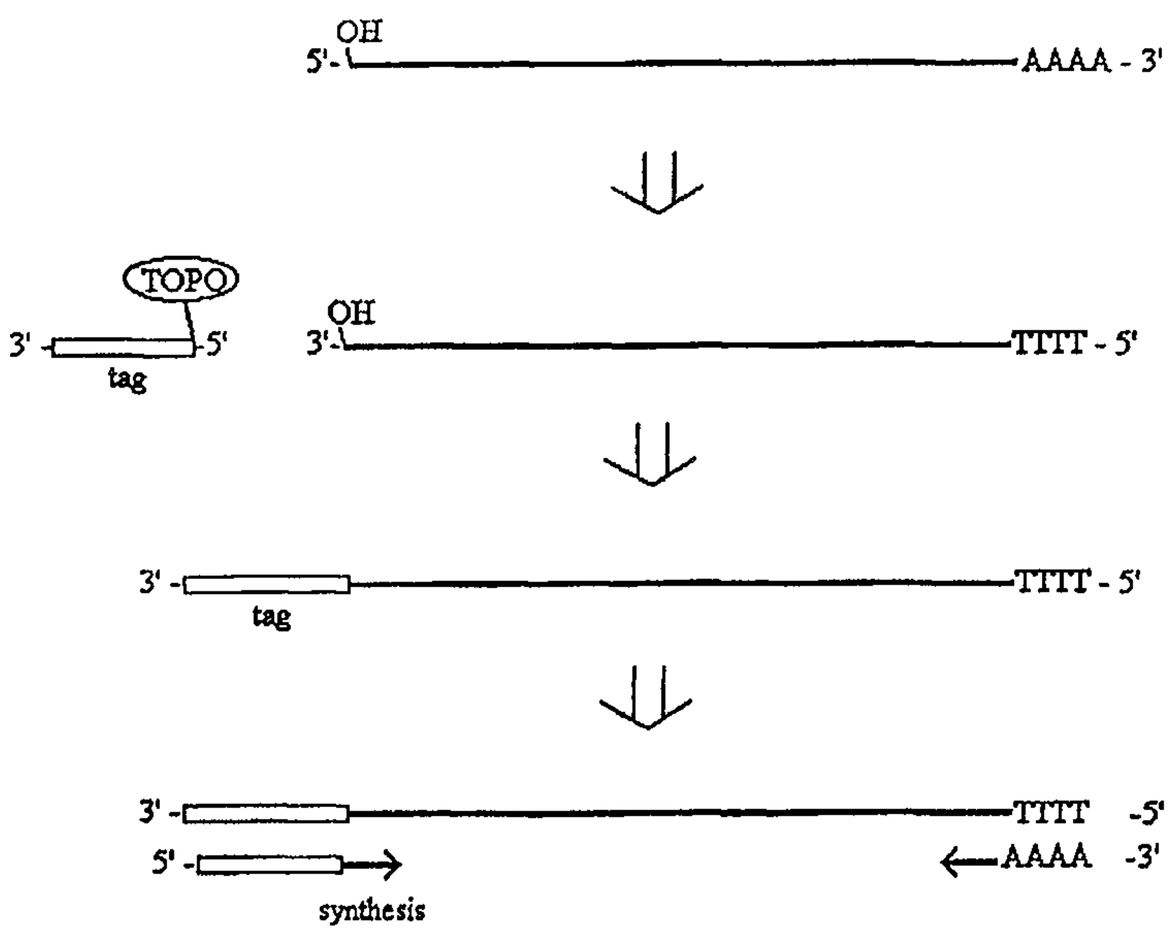
(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: HEYMAN, John, A. [US/US]; 6909 Quail Place, Carlsbad, CA 92008 (US). COMISKEY, John, D. [US/US]; 3764 Adams Street, Carlsbad, CA 92008 (US). MARCIL, Robert [US/US]; 1600 Faraday Avenue, Carlsbad, CA 92008 (US). PHELAN, Dorothy [IE/US]; 2320 Via Santos, Apartment S, Carlsbad, CA 92008 (US).

(74) Agent: REITER, Stephen, E.; Gray Cary Ware & Freidenrich LLP, Suite 1600, 4365 Executive Drive, San Diego, CA 92121 (US).

(81) Designated States: AE, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CR, CU, CZ, DE, DK, DM, DZ, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW, ARIPO patent (GH, GM, KE, LS, MW, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published
With international search report.
Before the expiration of the time limit for amending the claims and to be republished in the event of the receipt of amendments.

(54) Title: METHODS OF OBTAINING FULL–LENGTH NUCLEIC ACID SEQUENCES USING *E. COLI* TOPOISOMERASE III AND ITS HOMOLOGS



(57) Abstract

The invention disclosed herein comprises a method of obtaining full–length coding cDNA. The invention method comprises using isolated full–length mRNA to synthesize first strand cDNA, attaching a non–native tag to the cDNA and using the tagged first strand to synthesize second strand cDNA. The final cDNA can additionally be amplified and inserted into an expression vector. Also disclosed are isolated full–length coding cDNAs prepared using the method of the invention, as well as nucleic acid constructs containing the full–length coding cDNAs.

METHODS OF OBTAINING FULL-LENGTH NUCLEIC ACID SEQUENCES USING *E. COLI* TOPOISOMERASE III AND ITS HOMOLOGS

Field of the Invention

The invention disclosed herein relates to the fields of molecular biology and genomics and methods useful therein. More specifically, the invention relates to methods useful for the cloning and characterization of nucleic acid sequences encoding a full-length protein.

Background of the Invention

The fields of biology, agriculture and medicine have been significantly advanced by the ability to isolate DNA sequences that encode a particular protein, determine its sequence and eventually cause this DNA sequence to be expressed in a non-native host. This process has allowed potentially useful proteins to be more readily studied and has, furthermore, made them available as therapeutic molecules for the treatment of previously intractable disorders or as tools for the discovery of new therapeutics.

Recognizing the enormous value in identifying the DNA sequences that encode proteins, several groups have undertaken to sequence the entire genome of various organisms, including mice, yeast, several types of bacteria and humans. In general, genomic DNA is simply isolated, digested into fragments of a manageable size and sequenced using an automated sequencing system. Computerized searches of the resulting sequence data allow scientists to identify particular stretches of sequence that appear to encode a protein. This technique has proved useful when working with organisms having a relatively small genome. However, the genomes of complex organisms, such as humans, appear to contain a great deal of DNA sequence that does not encode proteins (introns and the like), thus blanket sequencing approaches may be a relatively inefficient means for identifying previously unknown, potentially useful proteins.

Techniques can be used to circumvent the issue of sequencing more DNA than one needs to identify new proteins. Typically, the relatively non-abundant messenger

RNA (mRNA) is isolated from a population of cells by contacting total isolated RNA with a poly-T sequence bound to a solid support under conditions where the 3' poly-A tail of the mRNA will bind to the poly-T sequence. Unbound molecules are washed away and the mRNA released with a low salt wash. cDNA is produced from the mRNA by 1) using reverse transcriptase to synthesize first strand DNA, 2) hydrolyzing the mRNA template by exposure to base, then 3) synthesizing second strand DNA with DNA polymerase. This results in a blunt-ended double-stranded sequence, which can then either be inserted into an expression vector directly or first tagged with a short linker nucleic acid having a restriction endonuclease recognition sequence contained therein.

This method, however, also has significant drawbacks. There is no assurance, for example, that the isolated mRNAs represent the full-length sequence encoding a protein. Furthermore, blunt-ended ligations require the use of T4 ligase, which is relatively slow acting, temperature sensitive and inefficient. Therefore, a need exists for a robust, straightforward method of obtaining full-length coding sequence that can subsequently be easily cloned into a vector for expression. The present invention addresses this and related needs.

Brief Description of the Invention

The present invention comprises a method of obtaining full-length coding sequences by, in part, employing the attachment of a non-native tag to a single-stranded cDNA product. In particular, the invention method comprises isolating full-length mRNA, using the isolated mRNA to synthesize first strand cDNA, attaching a non-native tag to the cDNA and using the tagged first strand to synthesize second strand cDNA. The final cDNA can additionally be amplified, using the polymerase chain reaction for example, and inserted into an expression vector.

A further aspect of the invention comprises isolated full-length coding cDNA prepared using the method of the invention, as well as nucleic acid constructs containing the full-length coding cDNA. The invention method is efficient and robust and provides cDNA products that can be of great benefit to those wishing to obtain

and study novel proteins and/or use said proteins as a means to develop new and useful substances.

Brief Description of the Figures

Figure 1 is a schematic representation showing attachment of a non-native tag to single-stranded cDNA using *E. coli* topoisomerase III.

Detailed Description of the Invention

In accordance with the present invention there are provided methods of obtaining full-length coding sequence from any organism. The invention methods comprise a series of steps: isolating full-length mRNA, synthesizing first strand cDNA using the mRNA isolated in the first step as a template, attaching a non-native tag to the 3'-end of the first strand cDNA and synthesizing second strand cDNA using the tagged first strand cDNA as a template. The invention method is unique in that it employs a novel method for the attachment of a non-native tag to a single-stranded DNA, in particular a tag comprising a sequence of nucleic acids having additional utility, such as, for example, providing for convenient cloning activities.

As used herein, the term "coding sequence" refers to the nucleic acids encoding the amino acid sequence of a protein, whereas a "gene sequence" refers to the entire nucleic acid sequence that is necessary for the synthesis of a functional polypeptide. A gene sequence could include, for example, a promoter in addition to a coding sequence. The term "cDNA" is used to refer to DNA that is complementary to mRNA, that is, unlike genomic DNA, it lacks introns.

The mRNA used in the method of the invention is typically derived from total RNA isolated from a population of cells. Techniques for isolation of total RNA are well known in the art. Exemplary techniques are described in Ausubel, et al, Short Protocols in Molecular Biology, 3rd ed. 1995, which is incorporated by reference herein in its entirety.

Total RNA is a mixed population of transfer RNA (tRNA) and mRNA from which the mRNA must be separated prior to initiating cDNA synthesis. The most

widely used technique exploits the poly-A tail found on the 5'-end of mRNA, as described above. This technique results in a population of mRNA sequences that contains a mixture of full-length and partial mRNAs, which then produces a population of cDNAs that will be a mixture of both full-length coding sequences and partial sequences. A more preferred method useful with eukaryotic cells isolates only full-length mRNA sequences utilizing the 5'-cap structure found on such sequences.

Most eukaryotic mRNAs are modified at their 5'-ends with a cap structure consisting of a 7-methylguanosine in a 5'-to-5' triphosphate linkage to the first transcribed nucleotide of the mRNA. This structure is added early during RNA transcription and is required as part of RNA biogenesis (see Sonenberg, N. Prog. Nucleic Acid Res. 35:173-207, 1988 for a review).

The cap structure can be employed to isolate full-length mRNA from other RNAs by several methods. In one method, the cap structure of RNA (total or mRNA previously isolated by oligo-dT techniques as described above) is modified by adding an affinity purification tag such as biotin, chitin binding domain, glutathione-S-transferase, and the like to the cap structure (Carnici, et al, DNA Res. 4(1):61-66, 1997). The affinity tagged capped mRNA can then be isolated from degraded mRNA or RNAs with poly-A tails that are not full-length. The affinity tagged mRNA is separated from untagged RNA using affinity purification, for example by contacting the tagged mRNA with an affinity purification material such as a solid support complexed with streptavidin, avidin, chitin, glutathione, and the like. Suitable solid supports include various column chromatography gels, such as sepharose, agarose, cellulose, and the like, and magnetic beads. Chromatography gels can be additionally modified with substances such as nickel.

Alternatively, unmodified capped mRNA can be separated from RNA species lacking a cap by several methods. For example, the capped mRNA can be contacted with a solid support complexed to, for example, phenylboronic acid or dihydroxylborate (see Theus and Liarakos, Biotechniques 9(5):610-612, 1990). The boronate ligand binds specifically and reversibly to vicinal diols such as the 2',3'-cis-diol of 7-methylguanosine ribose. The complex is disrupted by acidic pH and a

chelating agent such as EDTA, thereby releasing the mRNA from the affinity purification material.

Unmodified capped mRNAs may also be isolated using one or more cap-binding proteins bound to a solid support. Cap-binding proteins specifically recognize and bind to 7-methylguanosine. Several cap-binding proteins have been identified and isolated including eIF4E, eIF4A, eIF4G, eIF4B, and eIF4F, which exist in isomeric form in some species (see Haghighat and Sonenberg, JBC 272:21677-21680, 1997; van Heerden and Browning, J Biol Chem 269:17454-17457, 1994), and nCBP (see Ruud, et al, JBC 273(17):10325-10330, 1998). eIF4F is a multi-subunit complex composed of eIF4E, eIF4A and eIF4G.

Edery, et al reported a method for isolating capped-mRNA using eI4E bound to a solid support (Edery, et al, Mol. Cell. Biol. 15(6):3363-3371, 1995). eI4E alone, however, is much less efficient at interacting with capped mRNA compared to complexes comprising eIF4E and eIF4G or nCBP and eIF4G. Thus, a preferred method of isolating capped mRNA comprises using a eIF4E/eIF4G or nCBP/eIF4G complex bound to a solid support. Such complexes can be comprised of isolated, complexed individual proteins, or alternatively, a fusion protein comprising the cap-binding portions of the relevant cap-binding proteins.

Complexes suitable for use in the present invention can be produced in a variety of ways. Each of the components of the complex, generally two, can be isolated separately from a natural or recombinant source then either sequentially immobilized on a solid support or first mixed and allowed to form complexes in solution, which are subsequently bound to a solid support. A preferred method of forming complexes is to recombinantly produce them in the same cell line, so that complexes are formed within the cellular environment, purified, then bound to a solid support. Affinity purification tags, as described below, can be added to the recombinant proteins to aid in purification.

General procedures used to make fusion proteins are well known in the art. Fusion proteins that would be suitable for use in the practice of the present invention include a fusion between the C-terminus of full-length eI4E and the N-terminus of

full-length eI4G, with or without the addition of an affinity purification tag, or the N-terminus of full-length eI4E fused to the C-terminus of full-length eI4G, with or without an affinity purification tag.

Any cell type can serve as a source for mRNA to be used in practicing the method of the invention including both eukaryotic cells and prokaryotic cells, although only eukaryotic cells (such as plant, animal or insect cells) can be used as a source for mRNA isolated by the cap-binding procedure described above. Suitable animal cells include mammalian cells (human, rodent, non-human primate, goat, sheep, cow, and the like) and insect cells (moth, Drosophila, and the like). Cells in any stage of differentiation are suitable as a source for mRNA, as are cells in any particular activation state. Methods of extracting mRNA from different cell types are well known in the art (see, for example, Ausubel, et al, supra).

The isolated mRNA (either eukaryotic or prokaryotic) is used as a template for the synthesis of the first, or anti-sense, strand of the cDNA. Methods of first strand cDNA synthesis are well known in the art. Generally, a poly(dT)-containing oligonucleotide is used as a primer for synthesis by a reverse transcriptase enzyme. First strand synthesis techniques are described in detail in Ausubel, *supra*, and in the Examples below.

The mRNA/cDNA hybrids are treated with a substance that will degrade the mRNA, such as, for example, NaOH, yielding single-stranded first strand cDNA. In an alternative embodiment, prior to degradation of the entire mRNA strand, the mRNA/cDNA hybrid may first be treated with a substance that degrades single-stranded mRNA, such as, for example, RNAse. In this embodiment, any of the 5'-capped ends of any mRNA/cDNA hybrids wherein the cDNA is not full-length will be removed. After RNAse treatment, the treated hybrids are subjected to a cap protein-based affinity purification step as described above, removing any hybrids in which the first strand cDNA is not full-length. The isolated full-length RNAse-treated hybrids are then treated to degrade the mRNA, as described above.

A non-native tag sequence is next attached to the first strand cDNA. Scheele (US Patent 5,162,209, issued 11/10/92) describes a method wherein a homopolymeric

oligonucleotide tag is added to the 3' end of first strand cDNA using an enzyme such as terminal deoxyribonucleotidyl transferase, polyA polymerase, and the like. The resulting DNA either possesses a persistent homopolymeric oligonucleotide tag with no utility or, using the method described by Scheele, the tag is removed and the resulting cDNA is left with blunt ends, which reduces later cloning efficiency. An alternative method of adding a non-native tag sequence to single-stranded DNA uses T4 RNA ligase (see Tessier, et al., Anal Biochem 158:174-178, 1986; McCoy and Gumport, Biochemistry 19:635-642, 1980). This enzyme is, however, relatively slow and inefficient and does not work efficiently with any sequence (see Harada and Orgel, Proc. Natl. Acad. Sci. USA 90:1576-1579, 1993).

In contrast to the tags employed by Scheele, the non-native tags preferably employed in the invention method are nucleic acids which have some utility, such as comprising an enzyme recognition sequence that can be employed in subsequent cloning and/or insertion into a nucleic acid construct.

In accordance with the present invention, a non-native tag sequence is attached to the 3'-end of the first strand cDNA by a type I topoisomerase such as *E. coli* topoisomerase III, yeast topoisomerase III, and the like (see DiGate and Marrrians, JBC 264:17924-17930, 1989; Kim and Wang, JBC 267:17178-17185, 1992). These enzymes play a key role in DNA metabolism. Type I topoisomerases recognize and bind to specific recognition sequences in single-stranded DNA, catalyze a strand break and strand passage event, then reseal the break. An enzyme-DNA adduct is formed at the point of DNA cleavage, with the enzyme covalently bound to the nucleotide 5' to the cleavage site.

A preferred enzyme for use in the invention method is *E. coli* topoisomerase III (topo III). Topo III is a type I topoisomerase which recognizes, binds to and cleaves the single-stranded sequence 5'-GCAACTT-3' (SEQ. ID NO:1)(Zhang, et al, J. Biol. Chem. 270(40):23700-23705, 1995). A homolog, the traE protein of plasmid RP4, has been described by Li, et al., J. Biol. Chem 272:19582-19587, 1997. A DNA-protein adduct is formed with the enzyme covalently binding to the 5'-thymidine residue, with cleavage occurring between the two thymidines.

The non-native tag can be prepared by incubating a single-stranded nucleic acid molecule with topo III under conditions suitable for the enzyme's activity (discussed in greater detail in the Examples below). The nucleic acid will contain the topo III recognition sequence and generally a second sequence that may be used for priming second strand synthesis and/or in subsequent cloning activities, such as, for example, a recognition sequence for a restriction endonuclease or a recombinase which recognizes and acts on double-stranded DNA sequences such as Cre, Flp or vaccinia topoisomerase. Methods of cloning using vaccinia topoisomerase, for example, are described in U.S. Patent No. 5,766,891 (Shuman, issued June 16, 1998) which is incorporated by reference herein in its entirety.

The enzyme will bind to and cleave the nucleic acid molecule at the enzyme's recognition site and will, in the case of *E. coli* topoisomerase III, become covalently linked to the DNA 3' to the cleavage site via an enzyme-bridged phosphotyrosine linkage. This adduct can be isolated from cleavage products by well known treatment methods, for example treatment with detergents to temporarily inactivate the enzyme, followed by purification of the complex by standard chromatography or electrophoresis techniques, both of which are well known in the art. Alternatively, the tag-formation reaction can be performed while one end of the uncleaved molecule is immobilized and the clipped pieces removed by continual electrophoresis within an electric field.

In another alternative, the pre-cleaved nucleic acid molecule can be designed to contain a bridging phosphorothioate between the two thymidines of the GCAACTT cleavage/recognition sequence. When cleaved, the clipped piece will contain a 3'-SH instead of a 3'-OH, preventing religation (see Burgin, et al, Nucleic Acids Res 23:2973-2979, 1995).

In an additional embodiment, the DNA "tag" can be a nucleic acid vector, rather than a shorter nucleic acid molecule. Suitable vectors can be prepared as follows. A nucleic acid molecule suitable for attachment to a vector can be created by synthesizing complementary single strands of DNA which when annealed leave a single-stranded overhang containing a topoisomerase III recognition/cleavage

sequence. The double stranded end of the attachment molecule can either be blunt or contain a restriction endonuclease cutting sequence. A vector (as described below) can be linearized using a restriction endonuclease in such a way that the 5' end will match that of the attachment molecule. The attachment molecule is similarly treated, then ligated to the vector, forming a vector with a single stranded overhang containing the topoisomerase III recognition site. The modified vector is treated as described above to create an enzyme:vector adduct.

The topo III:tag adduct is incubated with the isolated first strand cDNA under conditions such that the enzyme will catalyze the joining reaction between the 3'-OH of the cDNA and the 5' end of the tag. Suitable conditions are discussed in greater detail in the Examples below.

Once the first strand cDNA and tag have been joined, the tagged cDNA is used as a template to synthesize second strand cDNA using the tagged first strand as a template. Methods of producing second strand cDNA are well known in the art (see, Ausubel, supra). One particularly useful method is the polymerase chain reaction (PCR), a technique well known in the art. The use of PCR has the advantage of amplifying the number of copies of any particular DNA sequence in addition to the creation of double-stranded DNAs. Nucleic acids complementary to the non-native tag and either a poly-dT or a nucleic acid complementary to the 5' end of the DNA can be used as primers for strand extension. If the first strand DNA is joined to a vector, DNA polymerase can be used to fill in the single stranded portions, creating blunt ends which can be ligated with T4 DNA ligase to create circular and transformable DNA according to techniques well known in the art.

In general, the procedure described above will efficiently produce full-length protein-encoding nucleic acids. The proteins encoded thereby can then be produced by inserting the newly cloned nucleic acid into an expression vector, which is subsequently transfected into a host cell for protein expression.

The amplified double-stranded DNAs can be isolated from the other components of the amplification reaction mixture prior to insertion into an expression vector. This purification can be accomplished using a variety of methodologies such

as column chromatography, gel electrophoresis, and the like. One method of purification utilizes low-melt agarose gel electrophoresis. The reaction mixture is separated and visualized by ethidium bromide staining. DNA bands that represent a majority of the amplification products are cut away from the rest of the gel and placed into appropriate corresponding wells of a 96-well microtiter plate. These plugs are subsequently melted and the DNA contained therein utilized as cloning inserts. The use of gel electrophoresis has the advantage that the practitioner can simultaneously purify the desired nucleic acid and verify that the sequence is of a reasonable size, i.e., probably represents the entire desired coding sequence.

The purified coding sequence can then be inserted into an expression vector. A variety of expression vectors are suitable for use in the method of the invention, both for prokaryotic expression and eukaryotic expression. In general, the expression vector will have one or more of the following features: a promoter-enhancer, a selection marker, an origin of replication, an affinity purification tag, an inducible element, an epitope-tag, and the like.

Promoter-enhancers are DNA sequences to which RNA polymerase binds and initiates transcription. The promoter determines the polarity of the transcript by specifying which strand will be transcribed. Bacterial promoters consist of -35 and -10 (relative to the transcriptional start) consensus sequences which are bound by a specific sigma factor and RNA polymerase. Eukaryotic promoters are more complex. Most promoters utilized in expression vectors are transcribed by RNA polymerase II. General transcription factors (GTFs) first bind specific sequences near the start and then recruit the binding of RNA polymerase II. In addition to these minimal promoter elements, small sequence elements are recognized specifically by modular DNA-binding/trans-activating proteins (e.g. AP-1, SP-1) which regulate the activity of a given promoter. Viral promoters serve the same function as bacterial or eukaryotic promoters and either provide a specific RNA polymerase in trans (bacteriophage T7) or recruit cellular factors and RNA polymerase (SV40, RSV, CMV). Viral promoters are preferred as they are generally particularly strong promoters.

Promoters may be, furthermore, either constitutive or, more preferably, regulatable (i.e., inducible or derepressible). Inducible elements are DNA sequence elements which act in conjunction with promoters and bind either repressors (e.g. lacO/LAC Iq repressor system in *E. coli*) or inducers (e.g. gal1/GAL4 inducer system in yeast). In either case, transcription is virtually "shut off" until the promoter is derepressed or induced, at which point transcription is "turned-on".

Examples of constitutive promoters include the int promoter of bacteriophage λ, the bla promoter of the β-lactamase gene sequence of pBR322, the CAT promoter of the chloramphenicol acetyl transferase gene sequence of pPR325, and the like. Examples of inducible prokaryotic promoters include the major right and left promoters of bacteriophage (P_L and P_R), the trp, reca, lacZ, LacI, AraC and gal promoters of *E. coli*, the α-amylase (Ulmanen Ett at., J. Bacteriol. 162:176-182, 1985) and the sigma-28-specific promoters *of B. subtilis* (Gilman et al., Gene sequence 32:11-20, 1984), the promoters of the bacteriophages of *Bacillus* (Gryczan, In: The Molecular Biology of the Bacilli, Academic Press, Inc., NY (1982)), *Streptomyces* promoters (Ward et at., Mol. Gen. Genet. 203:468-478, 1986), and the like. Exemplary prokaryotic promoters are reviewed by Glick (J. Ind. Microtiot. 1:277-282, 1987); Cenatiempo (Biochimie 68:505-516, 1986); and Gottesman (Ann. Rev. Genet. 18:415-442, 1984).

Preferred eukaryotic promoters include, for example, the promoter of the mouse metallothionein I gene sequence (Hamer et al., J. Mol. Appl. Gen. 1:273-288, 1982); the TK promoter of Herpes virus (McKnight, Cell 31:355-365, 1982); the SV40 early promoter (Benoist et al., Nature (London) 290:304-310, 1981); the yeast gal4 gene sequence promoter (Johnston et al., Proc. Natl. Acad. Sci. (USA) 79:6971-6975, 1982); Silver et al., Proc. Natl. Acad. Sci. (USA) 81:5951-5955, 1984), the CMV promoter, the EF-1 promoter, ecdysone-responsive promoter, and the like.

Selection markers are valuable elements in expression vectors as they provide a means to allow only those cells which contain a vector to grow. Such markers are of two types: drug resistance and auxotrophic. A drug resistance marker enables

cells to detoxify an exogenously added drug that would otherwise kill the cell. Auxotrophic markers allow cells to synthesize an essential component (usually an amino acid) while grown in media which lacks that essential component.

Common selectable markers include those which encode resistance to antibiotics such as ampicillin, tetracycline, kannamycin, streptomycin, hygromycin, neomycin, Zeocin™, and the like. Selectable auxotrophic genes include, for example, hisD, which allows growth in histidine free media in the presence of histidinol.

A preferred selectable marker for use in yeast expression systems is URA3. Laboratory yeast strains carrying mutations in the gene which encodes orotidine-5'-phosphate decarboxylase, an enzyme essential for uracil biosynthesis, are unable to grow in the absence of exogenous uracil. A copy of the wild-type gene (ura4+ in *S. pombe* and URA3 in *S. cerevisiae*) will complement this defect in trans.

A further element useful in an expression vector is an origin of replication. Replication origins are unique DNA segments that contain multiple short repeated sequences that are recognized by multimeric origin-binding proteins and which play a key role in assembling DNA replication enzymes at the origin site. Suitable origins of replication for use in expression vectors employed herein include *E. coli oriC*, 2μ and ARS (both useful in yeast systems), ColE1, sf1, SV40 (useful in mammalian systems), and the like.

Additional elements that can be included in an expression vector employed in the invention method are sequences encoding affinity purification tags or epitope tags. Affinity purification tags are generally short peptides that can interact with a binding partner immobilized on a solid support. Synthetic DNAs encoding multiple consecutive single amino acids, such as histidine, when fused to the expressed protein, may be used for one-step purification of the recombinant protein by high affinity binding to a resin column, such as nickel sepharose. An endopeptidase recognition sequence is engineered between the polyamino acid tag and the protein of interest to allow subsequent removal of the leader peptide by digestion with Enterokinase. Sequences encoding peptides such as the chitin binding domain (which binds to chitin), glutathione-S-transferase (which binds to glutathione), Thio-Patch

(LaVallie, et al. Bio/Technology 11:187-193, 1993) (which binds to metal-chelating resins), and the like can also be used for facilitating purification of the protein of interest. The affinity purification tag can be separated from the protein of interest by methods well known in the art, including the use of inteins (protein self-splicing elements, Chong, et al, Gene 192:271-281, 1997).

Epitope tags are short peptides that are recognized by epitope specific antibodies. A fusion protein comprising a recombinant protein and an epitope tag can be simply and easily purified using an antibody bound to a chromatography resin. The presence of the epitope tag furthermore allows the recombinant protein to be detected in subsequent assays, such as Western blots, without having to produce an antibody specific for the recombinant protein itself. Examples of commonly used epitope tags include V5, glutathione-S-transferase (GST), hemaglutinin (HA), the peptide Phe-His-His-Thr-Thr (SEQ ID NO:2), chitin binding domain, and the like.

A further useful element in an expression vector is a multiple cloning site or polylinker. Synthetic DNA encoding a series of restriction endonuclease recognition sites is inserted into a plasmid vector downstream of the promoter element. These sites are engineered for convenient cloning of DNA into the vector at a specific position.

The foregoing elements can be combined to produce expression vectors useful in expression of the coding sequences created using the invention method. Suitable prokaryotic vectors include plasmids such as those capable of replication in *E. coli* (for example, pBR322, ColEl, pSC101, PACYC 184, itVX, pRSET, pBAD (Invitrogen, Carlsbad, CA), and the like). Such plasmids are disclosed by Sambrook (cf. "Molecular Cloning: A Laboratory Manual", second edition, edited by Sambrook, Fritsch, & Maniatis, Cold Spring Harbor Laboratory, (1989) the relevant sections of which are incorporated by reference herein). *Bacillus* plasmids include pCl94, pC221, pTl27, and the like, and are disclosed by Gryczan (In: The Molecular Biology of the Bacilli, Academic Press, NY (1982), pp. 307-329). Suitable *Streptomyces* plasmids include plJlOl (Kendall et al., J. Bacteriol. 169:4177-4183,1987), and *Streptomyces* bacteriophages such as ϕC31 (Chater et al., In: Sixth

International Symposium on Actinomycetales Biology, Akademiai Kaido, Budapest, Hungary (1986), pp. 45-54). *Pseudomonas* plasmids are reviewed by John et al. (Rev. Infect. Dis. 8:693-704, 1986), and Izaki (Jpn. J. Bacteriol. 33:729-742, 1978).

Suitable eukaryotic plasmids include, for example, BPV, vaccinia, SV40, 2-micron circle, pCDNA3.1, pCDNA3.1/GS, pYES2/GS, pMT, pIND, pIND(Sp1), pVgRXR (Invitrogen), and the like, or their derivatives. Such plasmids are well known in the art (Botstein et al., Miami Wntr. Symp. 19:265-274, 1982; Broach, In: "The Molecular Biology of the Yeast *Saccharomyces*: Life Cycle and Inheritance", Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, p. 445-470, 1981; Broach, Cell 28:203-204, 1982; Dilon et at., J. Clin. Hematol. Oncol. 10:39-48, 1980; Maniatis, In: Cell Biology: A Comprehensive Treatise, Vol. 3, Gene Sequence Expression, Academic Press, NY, pp. 563-608,1980).

The coding sequences can be inserted into an expression vector using any of a variety of techniques. For example, if the non-native tag added to the cDNA contains a restriction endonuclease recognition sequence, the coding sequence can be ligated into a compatible sequence existing in the desired expression vector using standard techniques. A particularly preferred method is the use of the vaccinia topoisomerase cloning system described in US Patent No. 5,766,891, issued June 16, 1998 to S. Shuman and available as a cloning kit from Invitrogen, Corp., Carlsbad, CA.

Prokaryotic hosts are, generally, very efficient and convenient for the production of recombinant proteins and are, therefore, one type of preferred expression system. Prokaryotes most frequently are represented by various strains of *E. coli*. However, other organisms may also be used, including other bacterial strains. The prokaryotic host selected for use herein must be compatible with the replicon and control sequences in the expression plasmid. Recognized prokaryotic hosts include bacteria such as *E. coli* and those from genera such as *Bacillus, Streptomyces, Pseudomonas, Salmonella, Serratia*, and the like. However, under such conditions, the polypeptide will not be glycosylated.

Suitable hosts may often include eukaryotic cells. Preferred eukaryotic hosts include, for example, yeast, fungi, insect cells, and mammalian cells either *in vivo*, or

in tissue culture. Mammalian cells which may be useful as hosts include HeLa cells, cells of fibroblast origin such as VERO, 3T3 or CHOKl, HEK 293 cells or cells of lymphoid origin (such as 32D cells), and their derivatives. Preferred mammalian host cells include non-adherent cells such as CHO, 32D, and the like. Preferred yeast host cells include *S. pombe*, *S. cerevisiae* (such as INVSc1), and the like.

In addition, plant cells are also available as hosts, and control sequences compatible with plant cells are available, such as the cauliflower mosaic virus 35S and 19S, nopaline synthase promoter and polyadenylation signal sequences, and the like. Another preferred host is an insect cell, for example the *Drosophila* larvae. Using insect cells as hosts, the *Drosophila* alcohol dehydrogenase promoter can be used, (Rubin, Science 240:1453-1459, 1988). Alternatively, baculovirus vectors can be engineered to express large amounts of peptide encoded by a desired gene sequence in insect cells (Jasny, Science 238:1653, 1987); Miller et al., In: Genetic Engineering (1986), Setlow, J.K., et al., eds., Plenum, Vol. 8, pp. 277-297). The present invention also features the purified, isolated or enriched versions of the expressed coding sequence products produced by the methods described above.

The invention will now be described in greater detail by reference to the following non-limiting example.

Example

Full-length mRNA is isolated from a population of cells by first isolating total RNA using a commercially available kit, RNA Isolation Reagent (RNAwiz, Ambion) according to the manufacturer's instructions.

To prepare cap-binding protein affinity material, the 560 bp encoded wheat germ eIF-4E protein (the p86 subunit, Van Heerden and Browning, *supra*) is cloned into a vector such as pCR4 (Invitrogen), according to the manufacturer's instructions, to produce a Thio-Patch fusion protein. A chitin binding domain encoding sequence is also cloned upstream of the Thio-Patch site (pCR4), so that the resulting expressed eIF-4E protein contains both a chitin binding domain and Thio-Patch at the N-

terminus. A stop codon is introduced at the end of the insert so that the coding sequence will not read-through to the V5 epitope contained on the vector.

Vector transformed *E. coli* are grown and protein expression induced according to the manufacturer's instructions. Lysates are prepared by resuspending a liter of pelleted cells in Buffer A lysis solution (25 - 30 ml, Buffer A (10 mM tris, pH 7.5, 0.2 mM EDTA, 1 mM DTT, 10% glycerol, 100 mM KCl) plus 0.5% Triton X-100, 1 mM PMSF (phenylmethylsulfonyl fluoride)), repeated sonication, then centrifugation at 12,000 rpm for 15 minutes. The cleared lysate is removed from the pelleted cell debris and 1 mM PMSF added.

The protein is purified using m^7GTP Sepharose (Pharmacia). The resin is equilibrated with Buffer A, then added to the cleared lysate and stirred slowly overnight at 4°C. The resin is pelleted by centrifugation and the liquid removed. The resin is resuspended in Buffer A, washed by agitation at room temperature for 30 minutes, then re-pelleted. After removal of the liquid, the washed resin is resuspended in Buffer A and loaded into an empty column, allowing the Buffer A to drain through. The column is continuously washed with Buffer A until no protein is detected in the flow-through. The cap binding protein is eluted with 70 μM m^7GTP in Buffer A.

The isolated protein is then bound to a solid support such as Ni-ProBond Chitin Beads (New England BioLabs), used according to the manufacturer's instructions. Total RNA is added to the resin in a binding buffer containing 10 mM $KHPO_4$ (pH8), 100 mM KCl, 5% glycerol, 2mM EDTA, and 6 mM dithiothreitol (DTT) (added just before using) and supplemented with 1.3% polyvinyl alcohol (PVA). The RNA is mixed with the resin on an orbital shaker for approximately 1 hour at room temperature. The resin is then washed with binding buffer without PVA 3-4 times to remove unbound mRNA. After the final wash is removed, 600 μl of 10mM Tris (pH8) is added to the resin. The bound mRNA is eluted using an equal volume of phenol/chloroform, pH8. Following elution, the mRNA is precipitated with 0.4M NaOAc and an equal volume of isopropanol at -70°C for at least 30 minutes, then spun down in a microfuge to collect the RNA pellet. 20μg of mussel

glycogen is added as a carrier to aid in precipitation. The mRNA pellet is resuspended in 20μℓ RNAse-free water.

The isolated mRNA is converted to first-strand cDNA using a cDNA Cycle® Kit (Invitrogen) using the oligo dT primer provided and the protocols suggested by the manufacturer. The RNA is hydrolyzed by exposure to 0.2 N NaOH at 37° C for 30 minutes. The resulting single-stranded cDNA is tagged by the addition of previously prepared tag covalently bound to topoisomerase III (see below) in Tris pH8.0, 1mM Mg acetate, 0.1μg/ml BSA, NaC1 for 5 minutes at 37°C.

Topoisomerase III:DNA tags can be prepared by incubating a sequence such as, for example, 5'-NNNGCAACT*TCCCTATAGTGAGTCGTATTA-3' (SEQ ID NO:3) with *E. coli* topoisomerase III in 40 mM Hepes-KOH (pH 8.0 at 22°C), 0.1 mg/ml bovine serum albumin, and 1 mM magnesium acetate (pH 7.0) at 37°C for 3 - 5 minutes (where * marks the cleavage point). The reaction can be stopped by the addition of SDS or another mild detergent to a final concentration of 2%. The covalent enzyme:DNA adduct is isolated using column chromatography, for example using a size exclusion resin and HPLC.

PCR can be performed using the cDNA Cycle Kit according to the manufacturer's instructions and primers corresponding to the 3' end of the tag sequence and oligo-dT. The reaction cycles can be as follows: 2 minutes at 94°C, then 25 - 35 cycles (10 sec/cycle) at 94°C, 55°C and 72°C, followed by 5 minutes at 72°C. The resulting amplified cDNA can be inserted into a plasmid vector such as pCR®2 or pCR®2-Topo™ (Invitrogen, Carlsbad, CA) according to the manufacturer's instructions.

While the foregoing has been presented with reference to particular embodiments of the invention, it will be appreciated by those skilled in the art that changes in these embodiments may be made without departing from the principles and spirit of the invention, the scope of which is defined by the appended claims.

SEQUENCE LISTING

<110> Heyman, John A.
Comiskey, John D.
Marcil, Robert
Phelan, Dorothy

<120> METHODS OF OBTAINING FULL-LENGTH NUCLEIC ACID SEQUENCES USING E. COLI TOPOISOMERASE III AND ITS HOMOLOGS

<130> INVIT1220WO

<150> 60/125,126
<151> 1999-03-19

<160> 3

<170> FastSEQ for Windows Version 4.0

<210> 1
<211> 7
<212> DNA
<213> Escherichia coli

<400> 1
gcaactt 7

<210> 2
<211> 5
<212> PRT
<213> Artificial Sequence

<220>
<223> epitope-tag peptide

<400> 2
Phe His His Thr Thr
1 5

<210> 3
<211> 30
<212> DNA
<213> Artificial Sequence

<220>
<223> oligonucleotide for epitope tag preparation

<221> misc_feature
<222> (0)...(0)
<223> n = A, T, G, or C

<400> 3
nnngcaactt ccctatagtg agtcgtatta 30

That which is claimed is:

1. A method of producing full-length coding sequences, said method comprising:

(a) synthesizing first strand cDNA using isolated full-length mRNA from a population of cells as a template, thereby forming first strand cDNA/mRNA hybrid(s);

(b) denaturing the first strand cDNA/mRNA hybrid(s);

(c) attaching a non-native tag sequence to the 3' end of the first strand cDNA; and

(d) producing a full-length double-stranded cDNA by synthesizing second strand cDNA using the tagged first strand cDNA produced in step (c).

2. A method according to claim 1 wherein the mRNA is isolated employing an affinity purification material.

3. A method according to claim 2 wherein the affinity purification material comprises one or more cap-binding proteins bound to a solid surface.

4. A method according to claim 3 wherein the cap-binding protein(s) are selected from the group consisting of eIF4E, eIF4G, eIF4F, eIF4G, nCBP, and eIF4E:eIF4G fusion protein.

5. A method according to claim 2 wherein the mRNA to be isolated comprises a biotinylated cap structure.

6. A method according to claim 5 wherein the affinity purification material is a streptavidin or avidin-complexed solid support.

7. A method according to claim 1 wherein the mRNA is de-capped and de-phosphorylated after isolation.

8. A method according to claim 1 wherein the tag sequence comprises a recognition site for a site-specific recombinase.

9. A method according to claim 8 wherein the tag sequence further comprises a recognition site for a site-specific restriction endonuclease.

10. A method according to claim 1 wherein the tag sequence is attached by a site-specific recombinase capable of recognizing and acting on single stranded DNA.

11. A method according to claim 10 wherein the site-specific recombinase is *E. coli* topoisomerase III.

12. A method according to claim 1 further comprising amplifying the cDNA during or after step (d).

13. A method according to claim 12 further comprising inserting the amplified cDNA into an expression vector.

14. A method according to claim 1 further comprising treating the first strand cDNA/mRNA hybrid(s) formed in step (a) with a substance that degrades single stranded RNA; and isolating the undegraded hybrid(s) with an affinity purification material having affinity for capped mRNA prior to performing step (b).

15. A method according to claim 14 wherein the substance is RNAse I.

16. A method according to claim 14 wherein the affinity purification material comprises one or more cap-binding proteins bound to a solid support.

17. A method according to claim 14 wherein the mRNA component of the cDNA/mRNA hybrid comprises a biotinylated cap structure.

18. A method according to claim 17 wherein the affinity purification material is a streptavidin or avidin-complexed solid support.

19. A method according to claim 14 further comprising inserting the double stranded cDNA resulting from step (d) into an expression vector.

20. An isolated full-length coding sequence prepared according to the method of claim 1.

21. An expression vector comprising an isolated full-length coding sequence prepared according to the method of claim 1.

22. An expression vector according to claim 21 comprising one or more elements selected from: a promoter-enhancer, a selection marker encoding sequence, an origin of replication, an epitope-tag encoding sequence or an affinity purification-tag encoding sequence.

23. An expression vector according to claim 22 wherein the promoter-enhancer is the T7 promoter, gal1 promoter, metallothionein promoter, AraC promoter, or CMV promoter-enhancer.

24. An expression vector according to claim 22 wherein the selection marker encoding sequence encodes a protein which imparts antibiotic resistance to cells.

25. An expression vector according to claim 22 wherein the epitope-tag sequence encodes V5, the peptide Phe-His-His-Thr-Thr (SEQ ID NO:2), hemaglutinin, or glutathione-S-transferase.

26. An expression vector according to claim 22 wherein the affinity purification-tag sequence encodes a polyamino acid tag or a polypeptide.

27. An expression vector according to claim 26 wherein said polyamino acid tag is polyhistidine.

28. An expression vector according to claim 26 wherein said polypeptide is a chitin binding domain or glutathione-S-transferase.

29. An expression vector according to claim 26 wherein said polypeptide encoding sequence includes an intein encoding sequence.

30. An expression vector according to claim 21 wherein the expression vector is a eukaryotic expression vector or a prokaryotic expression vector.

31. An expression vector according to claim 30 wherein the eukaryotic expression vector is pYES2, pMT, pIND, or pcDNA3.1.

32. A method of obtaining full-length coding sequences comprising:

(a) contacting full-length mRNA, isolated from a population of cells by employing an affinity purification material, with reverse transcriptase, thereby synthesizing first strand cDNA and forming first strand cDNA/mRNA hybrids;

(b) treating the first strand cDNA/mRNA hybrids with a substance that degrades single stranded RNA;

(c) isolating undegraded hybrid(s) from degraded hybrids employing an affinity purification material having affinity for capped mRNA;

(d) denaturing the isolated cDNA/mRNA hybrids obtained from step (c) thereby producing single stranded cDNA and single stranded mRNA;

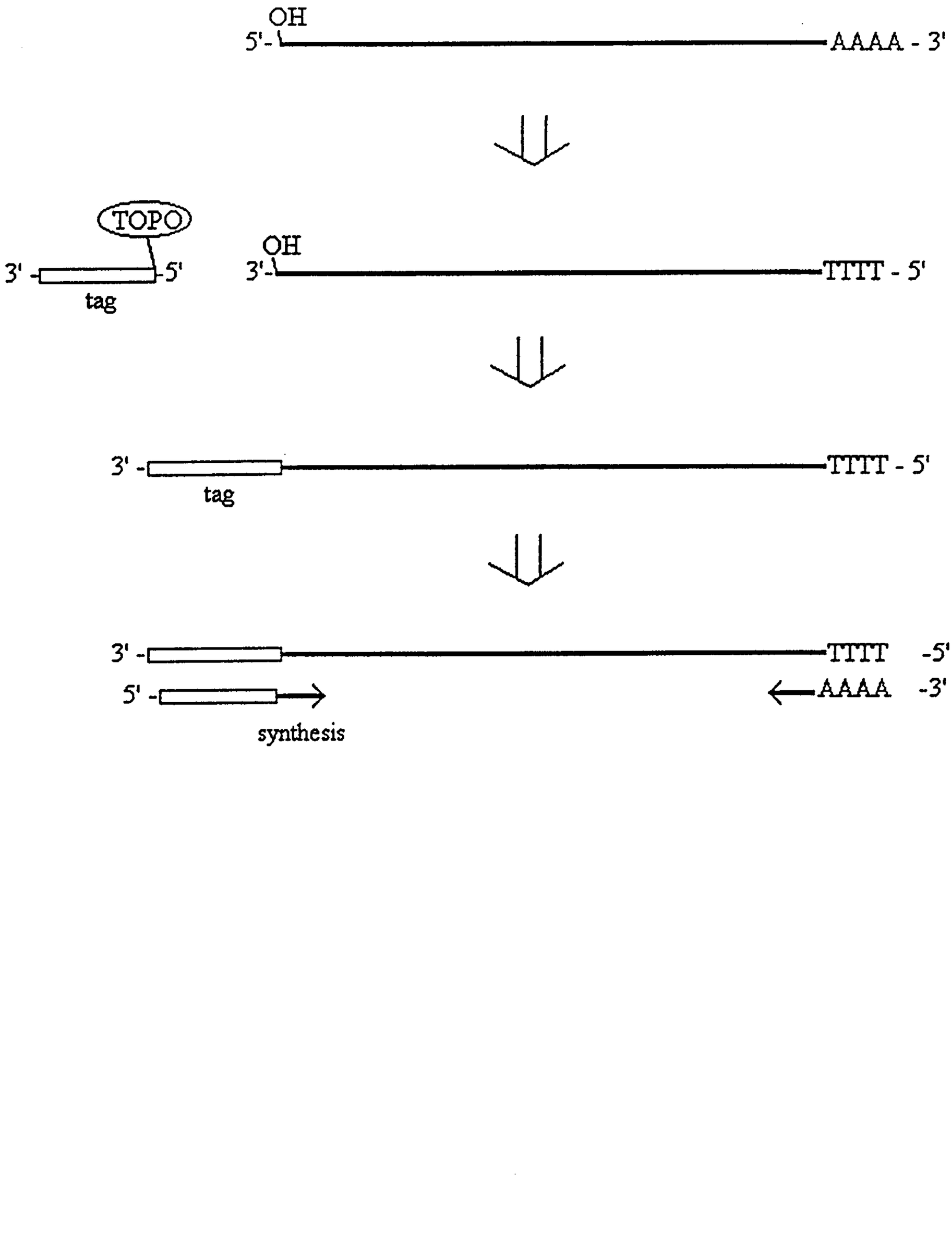
(e) attaching a non-native tag sequence to the single-stranded cDNA, wherein the tag sequence comprises a site-specific recombination sequence and is attached by *E. coli* topoisomerase III; and

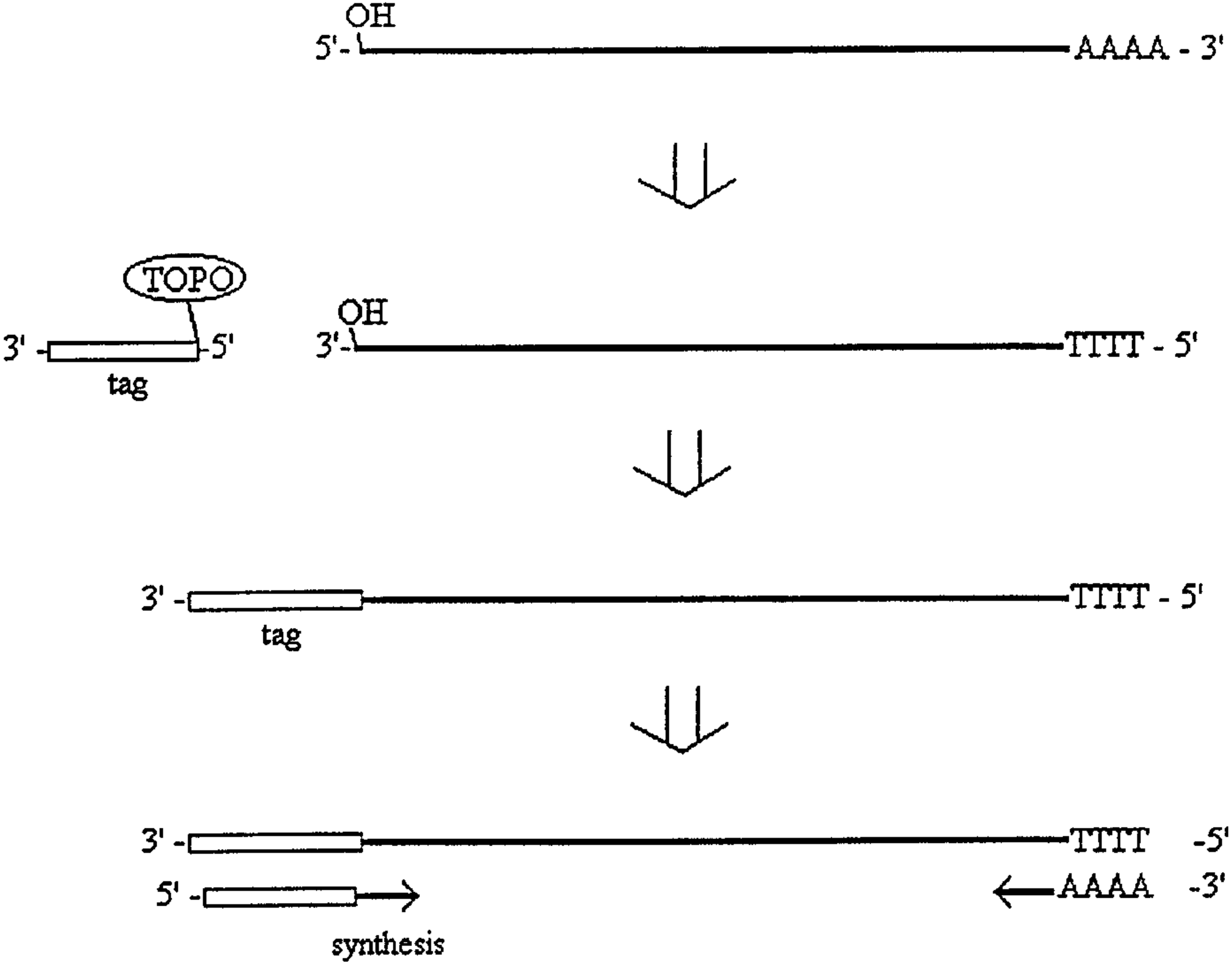
(f) synthesizing second strand cDNA using the tagged cDNA as a template and/or amplifying the cDNA, wherein the amplification primers comprise an anti-coding sequence of the tag sequence (5') and oligo-dT (3').

33. A method according to claim 32 further comprising inserting the cDNA obtained in step (f) into and expression vector.

34. A fusion protein comprising eIF4E and eIF4G.

Figure 1


OH
5'-
AAAA - 3'
TOPO
3'
5'
tag
OH
3'-
TTTT - 5'
3' -
tag
TTTT - 5'
3' -
TTTT -5'
5' -
AAAA -3'
synthesis


OH
5' -
AAAA - 3'
TOPO
3'
5'
tag
OH
3' -
TTTT - 5'
3' -
TTTT - 5'
tag
3' -
TTTT -5'
5' -
AAAA -3'
synthesis