(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
15 May 2003 (15.05.2003)

PCT

(10) International Publication Number
WO 03/040391 A2

(51) International Patent Classification[7]: C12Q

(21) International Application Number: PCT/GB02/05075

(22) International Filing Date:
8 November 2002 (08.11.2002)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
0126887.9 8 November 2001 (08.11.2001) GB

(71) Applicants *(for all designated States except US)*: **UNIVERSITY COLLEGE LONDON** [GB/GB]; Gower Street, London, Greater London WC1E 6BT (GB). **BIRKBECK COLLEGE** [GB/GB]; Malet Street, London, Greater London WC1E 7HX (GB). **THE INSTITUTE OF CANCER RESEARCH** [GB/GB]; 123 Old Brompton Road, London, Greater London SW7 3RP (GB).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **MCALISTER, Mark** [GB/GB]; 36 Ennerdale Drive, Congleton, Cheshire CW12 4FJ (GB). **SAVVA, Renos** [GB/GB]; 42B Raleigh Road, Harringay, London, Greater London N8 0HY (GB). **PEARL, Laurence** [GB/GB]; 211 Norwood Road, Herne Hill, London, Greater London SE24 9AG (GB). **PRODROMOU, Chrisostomos** [GB/GB]; 95 Fountains Crescent, Southgate, London, Greater London N14 6BD (GB). **DRISCOLL, Paul** [GB/GB]; 5 Talbot Road, Carshalton, Surrey SM5 3BP (GB).

(74) Agents: **KIDDLE, Simon, J.** et al.; Mewburn Ellis, York House, 23 Kingsway, London, Greater London WC2B 6HP (GB).

(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, 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, MZ, NO, NZ, OM, PH, PL, PT, RO, RU, SC, SD, SE, SG, SI, SK, SL, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, YU, ZA, ZM, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, SK, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:
— *without international search report and to be republished upon receipt of that report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

WO 03/040391 A2

(54) Title: METHOD FOR PRODUCING AND IDENTIFYING SOLUBLE PROTEIN DOMAINS

(57) Abstract: Methods for producing and identifying fragments of proteins, and more particularly to methods for generating and identifying soluble protein domains are disclosed based on a method for generating a library of nucleic acid fragments from nucleic acid encoding a desired polypeptide, and more especially a library of essentially, randomly sampled fragments of coding DNA sequence predominantly of defined size range and a method for selecting cloned gene fragments from the library that encode soluble protein domains.

Method for Producing and Identifying Soluble Protein Domains

Field of the Invention

The present invention relates to methods for producing and identifying fragments of proteins, and more particularly fragments which are soluble domains of a protein. The present invention further provides libraries of expression vectors and host cells comprising nucleic acid encoding the protein fragments and libraries of the protein fragments.

Background of the Invention

There are many large soluble, transmembrane and integral membrane multi-domain proteins of intense biomedical interest. These substances are by definition potential drug targets. Structural and functional analyses of these proteins will provide the basis for design of new strategies for therapeutic intervention in disease. High resolution structural study of proteins provides a basis for understanding biological and disease processes at molecular and atomic levels that is often necessary to support rational design or optimisation of new candidate drugs.

Biochemical and functional assays are used in drug discovery programs to identify compounds that interact with proteins in a manner that interferes with the biological function of the protein. These assays require large quantities of soluble protein to allow screening of thousands of compounds from chemical libraries. However, the production of sufficient quantities of these large proteins for detailed functional and structural studies is rarely feasible using existing methods. In the rare

cases where sufficient quantities of large multi-domain proteins can be produced, it is seldom possible to obtain the protein crystals that are prerequisite to structural study by X-ray crystallography or other techniques used in the art such as NMR. However, production of soluble fragments of these proteins may allow identification of regions of a protein that are responsible for the biological functions (or malfunctions), and facilitate detailed structural and functional analysis. Production of soluble protein fragments is therefore necessary to allow in-vitro biochemical and structural analyses of multi-domain proteins that cannot be obtained in sufficient quantities in intact form. However, little is known about the domain structure and organisation of many of these large proteins and bio-informatics approaches often do not provide a sufficient basis for rational identification of candidate domains. As a result, identification and expression of domains from many of these large proteins have proved refractory to the established, rational, recombinant protein engineering/expression strategies.

There are currently three main empirical approaches to identification of soluble protein domains: 1) bioinformatics and sequence analysis to estimate the location of domain boundaries of proteins based on sequence similarities with known proteins, 2) proteolytic fragmentation of the intact protein and identification of soluble fragments (REF), and 3) generation of "random" gene fragments, cloning to produce a gene fragment library and expression screening of the library to identify clones expressing soluble, folded protein fragments. Holistically these methods suffer from a number of weaknesses such as: a requirement for

quantities of the intact multi-domain protein for fragmentation that often cannot be obtained; failure to isolate gene fragments capable of producing soluble protein domains.

The most commonly used method for identification of minimal protein domains (domain-mapping) involves limited proteolysis of a target protein and identification of proteolytically resistant fragments by mass spectroscopy (e.g. Cohen, S. L. (1996)). This approach is based on the assumption that stable, folded domains are likely to be more resistant to proteolysis than unstructured regions of peptide sequence that are often found between domains. As this approach usually requires a reasonable quantity of highly purified, intact, soluble target protein derived from the native biological source, a large portion of human proteins of biomedical interest cannot be obtained in sufficient quantities. Protein samples are then enzymatically fragmented using various proteases. The molecular masses of the protein fragments generated are then measured by mass spectroscopy and the identity of the fragments may then be confirmed by further fragmentation (i.e. protein sequencing by MS). It is then assumed that protein fragments of around sixty or more amino acids residues in length represent stably folded domains since these portions of the protein appear to have greater resistance to degradation by proteases. This information is then used to design expression vectors for recombinant expression of the soluble domain candidates identified above.

In practice, there are several caveats with this approach that may result in failure to detect individual protein domains. The cleavage specificity of proteases is

limited to the peptide bond between certain amino-acid residue types (e.g. trypsin cleaves the peptide bond to the C-terminal side of basic residues). The position of protease cleavage sites is therefore not a function solely of structural context, but also of amino acid sequence context. Thus, if in practice the appropriate amino acid types are not found in a particular inter-domain peptide sequence, then the adjacent domains may not be separated and therefore the individual domains would not identified. In addition, steric hindrance may prevent protease-mediated cleavage of inter-domain peptide sequences that are short in length. Another major caveat of these approaches is that many domains comprise flexible loop regions that may be proteolytically sensitive resulting in cleavage within a domain (i.e. fail to detect the correct boundaries of a domain). Finally, a peptide sequence that corresponds to a soluble, folded proteolytic fragment may not necessarily be capable of autonomous folding and therefore recombinant over-expression of this particular peptide sequence may fail to produce soluble protein of tertiary structural integrity.

A DNA fragmentation based domain-mapping/identification method requires a protocol for generation of DNA fragments from an intact coding sequence in a manner that allows essentially random sampling of all possible fragments of appropriate size range (i.e. of a size capable of coding for a domain ~200-1500 nucleotides). In addition, the fragmentation protocol should ideally be generically reproducible, and must therefore be independent of differences in the properties of particular DNA targets, and produce fragments that are compatible with conventional methods for cloning of DNA

into vectors for protein expression. However, none of the existing DNA fragmentation methods fully meet requirements of random sampling, generic reproducibility, often displaying biased sampling and/or requiring optimisation of the method for particular target DNA properties such as DNA chain-length, and/or producing fragments that are incompatible with subsequent cloning applications. This is not surprising as many methods for fragmenting large DNA molecules have been developed for a wide variety of purposes other than protein domain identification.

A DNA fragmentation based domain-mapping/identification method requires a method for cloning of the DNA fragment mixture to produce a library of the gene fragments. A screening assay must then be used to identify clones that produce soluble folded protein fragments. A number of approaches have been developed for generation of libraries of different clones for a range of purposes including: large-scale DNA sequencing projects (e.g. shotgun cloning); selection of mutant proteins with particular enhanced functional properties (e.g. using gene-shuffling or random mutagenesis); and identification of epitopes for monoclonal antibodies by selection from a phage-display peptide library. Established library-based approaches to selection of protein variants or mutants have been recently adapted to identification of domains in large proteins including for example: a) cloning of DNA fragments into a bacteriophage surface-expression vector for expression as fusions with bacteriophage structural proteins (phage-display) using affinity selection as readout; b) cloning of DNA fragments into expression vectors to produce fusions with a reporter gene such as GFP or an antibiotic resistance gene, using

fluorescence and antibiotic resistance respectively as readout of recombinant protein solubility in vivo.

Phage display approaches involve enzymatic fragmentation of coding DNA and cloning of these fragments into a bacteriophage surface-expression vector to produce a phage display library of clones expressing different gene fragments on their surface. A method has been described involving shotgun cloning coupled with phage display mapping of functional domains of two streptococcal cell-surface proteins (Jacobson, et al., 1997). A phage-display library may be screened using a number of different approaches such as: target protein specific affinity selection and DNA sequencing of clones to identify the minimal fragment that retains binding affinity (e.g. Moriki et al., 1999); surface immobilisation of phage clones followed by limited proteolysis and washing to identify recombinant bacteriophage clones that are most resistant to proteolysis and are likely to display a fragment that has tertiary structure (Finucane et al., 1999). A limitation of affinity selection methods for screening of fragment libraries is a requirement for knowledge of the binding affinity(s) of the target protein, since this excludes the large number of proteins for which no specific binding or enzymatic activity has yet been established. Screening by limited proteolysis of phage particles adhered to a surface also suffers from the same caveats as other limited proteolysis approaches described above.

"Random PCR" has been used to generate fragments of target coding sequence for screening for soluble domains as fusions with green fluorescent protein (Kawasaki and Inagaki 2001). Caveats with this approach include:

"random PCR" is not truly random and will therefore not produce a complete library of all possible gene fragments of the appropriate size range; attachment of GFP to the expressed gene fragment may affect the folding and solubility of particular candidate domains resulting in both false negative and false positive results. An in vivo method for improvement of the solubility of proteins and protein domain constructs has been described involving mutagenesis of target proteins and production of fusions of target proteins with the antibiotic resistance gene chloramphenicol acetyl transferase and selection of clones with enhanced resistance to chloramphenicol (Maxwell et al., 1999). This method has not been used for domain identification. A caveat with this method is that there is only limited discrimination between soluble and insoluble proteins and the method does not select between folded and misfolded soluble fusions. An in vivo structural complementation based assay has been described involving fusions of the alpha fragment of beta-galactosidase with the C-terminus of target proteins so that if the fusion protein proves to be insoluble then interaction with the omega subunit will be prevented resulting in loss of beta-galactosidase activity (Wigley et al., 2001).

In summary, phage-display and fusion protein based methods have the common caveat that attachment of a reporter protein to a test protein is likely to influence the folding and solubility of the test protein in an unpredictable and target protein specific manner. In practice, existing DNA fragmentation approaches are not ideal for protein domain identification methods as none of these fully meet the requirements of random sampling, generic reproducibility and compatibility with subsequent

cloning applications. In addition, all existing methods for domain identification including limited proteolysis, gene fragmentation based methods such as phage display and fusion protein based screening methods all have serious limitations. These undoubtedly lead to failure to detect some protein domains and failure to identify the domains or regions of protein that are responsible for biological activities that could become the new targets for therapeutic intervention and drug development.

Summary of the Invention

Broadly, the present invention relates to methods for producing and identifying fragments of proteins, and more particularly to methods for generating and identifying soluble protein domains. In preferred aspects, the present invention is based on two innovative methods: 1) one relates to a method for generating a library of nucleic acid fragments from nucleic acid encoding a desired polypeptide, and more especially a library of essentially, randomly sampled fragments of coding DNA sequence predominantly of defined size range; and 2) a second relates to a method for selecting cloned gene fragments from the library that encode soluble protein domains.

In preferred embodiments, the present invention provides a holistic empirical method for the preparation and identification of regions of protein sequence that correspond to minimal domains or larger soluble fragments (e.g. several domains) and also permits production of these fragments in a form that is compatible with the structural and functional analyses identified above.

Accordingly, in a first aspect, the present invention provides a method for producing a library of nucleic acid fragments, the nucleic acid fragments encoding one or more portions of a polypeptide, the method comprising:

amplifying a nucleic acid sequence encoding the polypeptide in the presence of a non-native nucleotide so that the non-native nucleotide is incorporated into an amplified product nucleic acid sequence at a frequency related to the relative amounts of the non-native nucleotide and its corresponding native nucleotide, if present;

contacting the product nucleic acid sequence with one or more reagents capable of recognising the presence of the non-native nucleotide and cleaving the product nucleic acid sequence or excising the non-native nucleotide, thereby producing nucleic acid sequences encoding fragments of the polypeptide.

In a further aspect, the present invention provides a library of nucleic acid sequences encoding fragments of the polypeptide produced by the methods described herein.

In the present invention, "a non-native nucleotide" is a deoxynucleotide other than deoxyadenine (dA), deoxythymidine (dT), deoxycytosine (dC) or deoxyguanine (dG) that can replace the corresponding native nucleotide and is recognisable by the reagent used to cleave the product nucleic acid sequence or excise the non-native nucleotide from the product nucleic acid sequence. Preferably, the non-native nucleotides are neutral in terms of coding and are non-mutagenic. Examples of non-native nucleotides include uracil which can be used to replace thymidine and 3-methyl adenine which can be used to replace adenine.

Preferably, the amplification of the nucleic acid sequence is carried out using PCR using a non-native deoxynucleotide, either alone or in a mixture of the non-native and native nucleotide.

The starting nucleic acid sequence employed in the method may be a nucleic acid sequence encoding one or more polypeptide(s). In other embodiments, the starting nucleic acid comprises a cDNA or RNA library, or genomic DNA.

Preferably, the method comprises the further step of ligating the nucleic acid sequences encoding fragments of the desired polypeptide sequence into expression vector(s) to provide a library of expression vectors, and the optional further step of transforming host cells with the expression vectors to produce a library of host cells capable of expressing fragments (domains) of the polypeptide.

The method for generating random gene fragments involves random incorporation of a non-native nucleotide into the product nucleic acid sequence, in place of a native nucleotide, at a frequency that is preferably determined by the molar ratio of non-native to native nucleotide used in preparation of the coding sequence. The amplified nucleic acid product is then preferably contacted with a reagent capable of recognising and cleaving the sequence at the non-native nucleotide, for example by using an enzyme such as a DNA glycosylase or endonuclease, which can recognise the presence of the non-native nucleotide and cleave the nucleic acid sequence at or around the non-native nucleic acid sequence. A preferred protocol employs enzyme(s), β-

elimination and temperature changes in order to generate DNA fragments derived by essentially unbiased sampling and predominantly of defined size range. The method of the present invention is particularly advantageous as it allows the production of nucleic acid fragments of a size which encode protein domains of the polypeptide, e.g. preferably between 100 and 1500 nucleotides, more preferably between 200 and 1200 nucleotides, and most preferably between 300 and 1000 nucleotides in length, and is capable of fine sampling of the nucleic acid encoding the polypeptide, producing fragments on average every second nucleotide. In order to allow generic application for library sampling of any polypeptide it may be advantageous to re-code certain nucleotide sequences to contain more incorporation sites for the non-native nucleotide, up to the limits imposed by the constraints of the genetic code.

Optionally, the nucleic acid fragments may then be further amplified to produce nucleic acid fragments for further uses. Additionally or alternatively, the nucleic acid fragments may be exposed to enzymes that mediate attachment of the fragments to other DNA molecules, such as an expression vector, comprising sequences responsible for control of transcription and translation of the gene fragments and optionally sequence encoding affinity tag peptide sequences and optionally sequence for replication of the derived DNA constructs in host cells to produce gene fragment expression constructs.

Thus, in a preferred embodiment, the present invention provides a method of producing fragments of a desired polypeptide, the method comprising expressing the nucleic acid sequences encoding fragments of the desired

polypeptide and optionally isolating the polypeptide fragments thus produced. Preferably, the polypeptide fragments are expressed as fusions with an affinity tag, so that they can be purified by affinity chromatography. Preferably, peptide based affinity tags will be less than 25 amino acid residues long, and more preferably less than 15 residues long. Preferred affinity tags have minimal effect on the solubility, stability and/or aggregation state of the attached protein fragment. The use of C-terminal affinity tags is preferred as this permits the selection of clones that express in-frame fragments of DNA, while DNA fragments which are out-of-frame would tend to terminate prior to the translation of the tag.

Examples of suitable affinity tags include polyhistidine (e.g. the hexa-His tags exemplified herein) which bind to metal ions such as Ni^{2+} or Co^{2+}, Flag or Glu epitopes which bind to anti-Flag antibodies, S-tags which bind to streptavidin, calmodulin binding peptide which binds to calmodulin in the presence of Ca^{2+}, and ribonuclease S which binds to aporibonuclease S. Examples of other affinity tags that can be used in accordance with the present invention will be apparent to those skilled in the art.

In a further aspect, the present invention provides a library, e.g. as produced by a method of described herein, which is:

(a) a library of nucleic acid fragments of a parent nucleic acid sequence, wherein the nucleic acid fragments have a size range as disclosed herein and are preferably sampled from the parent nucleic acid sequence on average about every second nucleotide; or

(b) a library of expression vectors which comprise a plurality of the nucleic acid fragments as set out in (a), wherein each fragment is ligated to a nucleic acid sequence encoding an affinity tag and optionally one or more further sequences to direct the expression of the nucleic acid fragment and the affinity tag; or,

(c) a library of host cells transformed with the expression vectors as defined in (b); or

(d) a library of polypeptide fragments produced by expressing the nucleic acid sequences, wherein each polypeptide is coupled to an affinity tag.

Preferably, this method makes use of non-native nucleotides, and in particular non-native nucleotide bases that can be randomly incorporated into the DNA duplex and then selectively excised to produce the nucleic acid fragments of the polypeptide. None of the current enzymatic methods reviewed above, that aim to produce DNA fragments of essentially random distribution with respect to the source DNA (e.g. DNAase 1 digestion), provide robust control of fragment size range or sampling of DNA in a manner fully independent of DNA secondary structure, or robust reproducibility. In contrast, the present method preferably provides fine sampling with cleavage every second nucleotide on average, robust control of fragment size range, rapid and facile execution, and robust reproducibility. The DNA produced by the method is also compatible with blunt ended and TA cloning methods for construction of expression vectors.

In a preferred embodiment, the present invention employs a DNA fragmentation method based upon an enzymatic fragmentation DNA base-excision repair pathway, (Savva, et al., 1995; Savva and Pearl, 1995; Panayotou, et al.,

1998; Barrett, et al., 1998; Barrett et al.,1999). This system initiates the removal of uracil the pro-mutagenic deamination product of cytosine from DNA by the sequential hydrolysis of the bond linking the base to the sugar, followed by cleavage of the sugar phosphate backbone at the abasic site by an apurinic/apyrimidinic endonuclease (APE). The initial reaction, catalysed by uracil-DNA glycosylase (UDG) is exquisitely specific for uracil, and proceeds with very high efficiency. Thus, exposure to UDG and APE enzymes produces a single-strand nick in a dsDNA molecule wherever a uracil occurs. Like the normal DNA component thymine (identical to 5-methyl-uracil), uracil forms stable Watson-Crick base pairs with adenine, and can be efficiently introduced into dsDNA by template-dependent DNA polymerase reactions, using Pol 1 family enzymes such as Taq in PCR reactions. The widely used archaeal DNA polymerases such as Pfu or Vent are inhibited by template strand uracil (Greagg et al., 1999) and are not suitable for this purpose. Incorporation of uracil opposite a template-strand adenine occurs with comparable efficiency to incorporation of thymine, and is unbiased by sequence context. Thus, the probability of uracil incorporation in the daughter strand opposite a template-strand adenine is purely a function of the ratio of TTP/dUTP present in the PCR reaction mix and independent of uracil incorporation in previous cycles. The product of an 'ideal' TTP/dUTP PCR reaction is a mixture of otherwise identical double-stranded DNA molecules in which each possible thymine in either strand has been replaced by uracil. PCR under these conditions is robust even for relatively large PCR products. When this reaction mixture is exposed to UDG and APE to completion, single-strand breaks are introduced at each position at which a uracil was incorporated. A typical

mammalian genome has a thymine content ≈ 25%, therefore double stranded DNA fragments are generated beginning and ending ≈ every 2nd base since cleavage may occur at uracil sites on both coding and non-coding strands.

Cleavage by APE leaves a deoxyribose phosphate moiety at the 3′ or 5′ side of the nick, depending on the specificity of the APE used. The deoxyribose phosphate moiety may then be removed by β-elimination, which is accelerated by mild bases such as spermine and elevated temperature (Bailly and Verly, 1989) to produce single nucleotide gaps in one strand of the duplex. In order to produce blunt-ended DNA fragments for cloning two alternative approaches may be used: 1) cleavage of the single-stranded DNA opposite the single-nucleotide gaps in the duplex DNA using S1-nuclease (Vogt, 1973)(Figure 1); 2) thermal denaturation of the duplex DNA and re-annealing of the DNA at reduced temperature and filling of 3′ recesses using a template dependent DNA polymerase, followed by removal of 3′ extensions using a single-strand specific exonuclease with 3′-5′ exonuclease activity such as Mung bean nuclease or a single-strand specific endonuclease such as S1-nuclease (Figure 2).

This DNA fragmentation method has several advantages over other possible methods. Firstly, given pure reagent enzymes, every enzymatic step can be allowed to go to completion, so that the size distribution of the fragments generated, is dictated solely by the TTP/dUTP ratio used in the original PCR reaction. This is in contrast to other enzymatic digestion approaches such as: cleavage by endonucleases (eg. DNAase I) that cleave both strands of duplex DNA, which fully degrade DNA to free nucleotides if the digestion is allowed to go to

completion. Computer simulations of the present method using a 5120 base pair gene suggest that a TTP/dUTP ratio of 100:1 will give even cover of the coding sequence, and good representation of fragments in the desired 'domain' size range (~300-1000 nucleotides). Secondly, all the procedures involved are enzymatic and therefore carried out under 'mild' conditions that will cause no other DNA damage, and are completely compatible with rapid efficient DNA purification methods such as ion-exchange and silica-based adsorption methods that may be used between subsequent steps. Thirdly, the products of these reactions are fully 'biological' and suitable for cloning into expression vectors by blunt-end ligation or TOPO-isomerase I-mediated ligation.

It would also be possible to employ a different non-native nucleotide and use a corresponding enzyme which is capable of recognising the non-native nucleotide in the amplified nucleic acid sequence and removing it from the amplified nucleic acid sequence or cleaving the sequence, thereby generating the fragments. One example is 3-methyladenine-DNA glycosylase from *E.coli* which is another monospecific DNA glycosylase that could also be used if deoxy-3-methyladenine (3-meA) mononucleotides are incorporated instead of deoxyadenine (both form base pairs with thymidine). This nucleotide could be generated by exposing deoxyadenine mononucleotides to the methylating agent methyl methanesulphonate (MMS) and re-purifying them.

In many circumstances, it will be desirable to generate 'ragged-terminus' libraries in which, for example, a domain such as an N-terminal domain is always present, but a wide range of C-termini are to be sampled. This

can be readily achieved using the method by performing two PCR steps and a thermal denaturation and annealing step: 1) amplification of the constant 5′-segment encoding the N-terminus in a TTP PCR reaction; 2) amplification of a 3′ segment that partially overlaps with the 5′ segment in a TTP/dUTP PCR reaction; 3) and then mixing the products of these two PCR reactions before thermal melting and re-annealing. A restriction endonuclease (RE) site, that generates a "sticky-ended" on cleavage, may be introduced into the 5′ extremity of the 5′-segment, so that the library of N-terminally constant but C-terminally ragged coding sequences can then be efficiently cloned into a vector cleaved the above RE and another with a second RE that generates a blunt end.

In a further aspect, the present invention provides a method of identifying soluble protein domains, the method comprising:

expressing a library of nucleic acid fragments to produce the protein domains encoded by the fragments, wherein the protein domains are expressed as fusions with an affinity tag; and

separating soluble proteins using the affinity tag.

Examples of affinity tags that can be employed in the present invention are provided above and many others will be apparent to the skilled person. The use of C-terminal affinity tags is preferred as this permits the selection of clones that express in-frame fragments of DNA, while DNA fragments which are out-of-frame would tend to terminate prior to the translation of the tag.

The method may comprise the additional step of

identifying soluble proteins which are domains of the polypeptide, e.g. share a binding or biological activity with the full length parent polypeptide.

Optionally, the method comprises making a library of soluble protein fragments or domain and contacting the fragments or domains with one or more candidate compounds to determine whether one or more of the candidate compounds binds to and/or modulates an activity of a protein fragment or domain present in the library. The candidate compounds may be small molecules or alternatively candidate polypeptide binding partners, e.g. the method can be used to investigate ligand-receptor binding, enzyme-substrate binding, antibody-antigen binding, protein-ligand binding or protein-nucleic acid binding. In still further embodiments, two or more libraries of soluble protein fragments or domains can be crossed to determine whether binding or modulation of activity occurs between members of the libraries. By way of example, in this embodiment of the invention, libraries of domains of two proteins can be made to determine which portions of those proteins are involved in binding and biological activity.

In this aspect of the present invention, the nucleic acid fragments is introduced into an expression vector(s) to produce a library of different DNA fragment expression constructs and protein expression is induced and the derived protein then treated in a novel approach that selectively removes insoluble and/or soluble misfolded and/or non-specifically aggregated protein fragments allowing selective detection and purification of the soluble folded unaggregated or specifically aggregated protein fragments.

The approach makes use of the observation that empirically the process of purification of affinity tagged (such as hexahistidine tagged) proteins by affinity chromatography (such as metal affinity chromatography) is strongly selective for soluble, folded proteins. Selection occurs in several stages in the purification method including: loss of insoluble protein at filtration or centrifugation steps; loss of weakly soluble, misfolded or non-specifically aggregated protein by precipitation or non-specific binding to various surfaces such as plastic and glass surfaces at all stages of purification; loss of misfolded or non-specifically aggregated protein by failure to adsorb to affinity media, and/or loss at washing steps. In our studies, affinity tags, such as the hexa-histidine tag, appear to display considerably lower accessibility to affinity chromatographic media when attached to misfolded, aggregated and/or insoluble target proteins, rather than to stably-folded, un-aggregated, soluble target proteins. This selectivity is likely to result in part from differences in the degree of steric hindrance of binding to affinity media, resulting from the properties of the target protein (e.g. soluble vs. insoluble, folded vs. misfolded, non-specifically aggregated vs. un-aggregated or specifically aggregated). In this novel method, the DNA fragment expression library is induced and screened for soluble protein expression on the basis of the selectivity of affinity purification media for binding of folded, soluble tagged proteins over misfolded, insoluble or aggregated tagged proteins.

In some embodiments, the blunt-ended DNA fragments may be operationally linked to DNA sequences such as an

expression vector, comprising sequences responsible for control of transcription and translation of the gene fragments and optionally sequence encoding affinity tag peptide sequences and optionally sequences for replication of the derived DNA constructs in host cells.

In some embodiments the library of blunt ended gene fragments are ligated into a suitable expression vector using conventional blunt-ended ligation methods. Alternatively, the blunt-ended gene fragments are cloned into a suitable expression vector. An inducible expression vector may be used such as those based on the pET series in which the restriction fragments can be inserted between the T7 promoter and start codon at the 5′ end, and stop-codons and transcription terminator at the 3′ end. Different versions of the vector may be constructed, to include an affinity tag (e.g. a His_6-tag) and an optional protease cleavage site at the N-terminus or C-terminus of the expressed fragment. A number of different vectors may be employed to provide start and stop codons in all three reading frames. The procedures described here are not limited to the use of the His_6-tag, and allow for the use of alternative tags and/or development of alternative short tags compatible with fluorescence or FRET-based protein detection strategies for example. The expression vectors constructed above constitute a gene fragment expression library. This library is then transfected into host cells and the transformed cells then spread on to selection media plates.

Several hundreds or thousands of individual colonies may then be picked from the selection media plates and transferred to multi-well growth plates containing

suitable growth medium. Several hundreds or thousands of clones may be analysed, so that all subsequent stages may be processed in parallel utilising multi-well formats implemented on a multi-well plate format liquid-handling robot. Plates are incubated at 15-37°C overnight, and aliquots transferred into a second plate for growth for 2-3 hours. Optionally, expression may be induced by addition of inducer molecules or temperature change, and cultures grown for a further period post-induction. Alternatively, a constitutive promoter system may be utilised. Cell-growth is monitored by optical density measurement. The cells are then lysed and then contacted with appropriate affinity chromatography media such as metal chelate media in conditions under which insoluble or soluble mis-folded protein molecules are removed by precipitation or adsorption onto surfaces, such that only soluble folded protein fragments are efficiently purified. The purified soluble protein fragments are analysed with respect to concentration and covalent structural integrity.

Preferably, the expressed proteins are released for separation under non-denaturing conditions, e.g. by enzymes, or non-denaturing detergents. Thus, host cells such as induced bacterial cells are lysed using lysozyme and non-denaturing detergents, and the lysates applied to a multi-channel filter system (e.g. Qiagen TurboFilter) that removes unbroken cells, cell debris and insoluble material. Alternatively, the lysates may be clarified by centrifugation. The clarified lysates containing the soluble contents of the induced cells are then purified in parallel in multiwell format by affinity chromatography (e.g. metal affinity chromatography) and assayed by anti-tag immunoblot or ELISA, SDS-PAGE and

mass spectrometry and other methods known to those skilled in the art. This combination of readouts guarantees high sensitivity (blot or ELISA), assessment of purity (SDS-PAGE) and validation of the molecular composition, in addition to quantifying the protein expression level. In an alternative configuration of this embodiment, multiple clones are individually picked from the selective media plate and then cultured together in selective liquid media and processed together at all subsequent steps in order to reduce the total number of parallel operations to be performed. The chances of any one fragment of the appropriate size range corresponding to a folded domain and therefore giving a positive readout is likely to be 0.01-1%. In this context, when a pool of clones gives a positive readout then each original clone present in the pool or subpool is reprocessed to identify which clone(s) produced the positive readout.

In a further alternative embodiment, all colonies from the selection media plates may be pooled and cultured in single vessel containing selective liquid media as described above with respect to temperature and induction of expression, before cell lysis and purification by affinity chromatography. In this embodiment, the purified protein mixture is then analysed as described above and is likely to be found to contain multiple soluble protein fragments, which can be identified by protein sequencing and/or by fragmentation mass spectroscopy. The coding DNA sequences corresponding to the protein fragments identified are then amplified by PCR and cloned into expression vectors using established methods known to those skilled in the art and used for large-scale preparation of the protein fragment. In this

context different versions of expression vectors may be constructed, to include an affinity tag (e.g. His_6-tag) and an optional protease cleavage site at the N-terminus or C-terminus of the expressed fragment.

Once clones that express soluble protein fragments have been identified these clones are then cultured on a larger scale with optional optimisation of expression, and processed as described above, before purification employing the affinity tag, e.g. employing chromatography media and methods well known to those skilled in the art. The purified soluble protein fragments are analysed with respect to concentration, covalent structural integrity, tertiary structural integrity and biological and/or enzymatic activity using methods well known to those in the art.

One embodiment of this method seeks to identify soluble fragments of an extracellular protein or extracellular domains of a transmembrane or integral membrane protein that are suitable for high-level expression and secretion in bacterial systems. In this embodiment, the library of nucleic acid fragments is cloned into an expression vector that fuses a bacterial periplasmic export signal (such as OmpA) and signal peptidase cleavage site to the N-terminus of the expressed protein fragment. An affinity tag can optionally be included following the signal peptidase site or at the C-terminus of the expressed protein fragment. Bacterial colonies expressing these protein fragments are treated with gentle osmotic shock to release proteins from the bacterial periplasmic space, with minimal release of proteins from the cytoplasm. The periplasmic contents and bathing culture medium are then filtered and contacted with affinity resins as in the basic

methodology. In this embodiment, only those protein fragments that were efficiently secreted into the periplasmic space, were proteolyticaly released from the signal peptide, and were soluble and unaggregated following secretion from the cells or after osmotic shock, are efficiently purified and will give strong anti-tag signals in immunoblot or ELISA assays.

A further embodiment of the method seeks to identify candidate surface proteins from bacteria, suitable for vaccine development. In this embodiment, the method for identification of soluble fragments suitable for high-level expression and secretion in bacterial systems described above, is applied to screening a DNA fragment library derived from part of, or an entire bacterial genome, generated by some type of DNA fragmentation method. DNA fragments from such a library will be cloned into the expression vector for periplasmic export, and colonies screened for expression of soluble tagged-protein fragments in culture medium and periplasmic extract. Those expressed protein fragments that give strong anti-tag signals, will be those that were efficiently secreted into the periplasmic space, were proteolyticaly released from the signal peptide, and were soluble and unaggregated. It is most likely that protein fragments that fulfil these criteria efficiently will derive from extracellular proteins, or from the extra-cellular domains of transmembrane or integral membrane proteins, encoded by the bacterial genome being screened. Such proteins would have a high likelihood of being visible to the immune system of an organism infected by the bacterium being screened, and would therefore be good candidates for vaccine development.

In a further variation, the method can be used to identify stable and soluble complexes formed between fragments of different proteins or between fragments of a single protein. In one embodiment of this variation, two or more DNA fragment libraries are co-expressed in the same bacterial cell, either from the same vector, or from different compatible vectors simultaneously present. The libraries are cloned into the expression vector or vectors as in the basic method, but so that sequences encoding different affinity 'tags' are attached to the fragments encoded by the different DNA libraries. As in the basic method, bacterial cells are lysed and filtered, and contacted with affinity media that is specific to the (primary) affinity tag attached to only one library to select for soluble, folded and unaggregated protein fragments. As in the basic method protein levels are assayed by ELISA or immunoblot, but using antibodies directed against the (secondary) affinity tag (or tags) attached to the other library (or libraries). Strong signals against a secondary tag, will indicate the presence of a fragment expressed from one library, that was efficiently transported by and formed a stable non-aggregated complex with a fragment from the primary library whose 'tag' was utilised for selection.

In a further aspect, the methods described herein may be combined to provide a method of producing a library of nucleic acid fragments, the nucleic acid fragments encoding one or more portions of a polypeptide, and identifying fragments encoding soluble protein domains, the method comprising:

amplifying a nucleic acid sequence encoding the polypeptide in the presence of a non-native nucleotide so that the non-native nucleotide is incorporated into the

amplified product nucleic acid sequence at a frequency related to the relative amounts of the non-native nucleotide and its corresponding native nucleotide, if present;

contacting the product nucleic acid sequence with one or more reagents capable of recognising the presence of the non-native nucleotide and cleaving the product nucleic acid sequence or excising the non-native nucleotide, thereby producing nucleic acid sequences encoding fragments of the polypeptide;

expressing a library of the nucleic acid fragments to produce the protein domains encoded by the fragments, wherein the protein domains are expressed as fusions with an affinity tag; and

separating soluble proteins using the affinity tag.

Embodiments of the present invention will now be described in more detail by way of example and not limitation with reference to the accompanying figures.

Brief Description of the Figures

Figure 1 shows a representation of the fragmentation of a single molecule of PCR product with a low level of dUTP incorporated. Since the position at which the dUTP is incorporated is different in different PCR product molecules, the position at which cleavage occurs is different and will therefore result in sampling of all possible positions in a particular coding sequence. A library of DNA fragments are therefore produced that sample all possible positions representing all possible fragments within a certain size range, that is determined by the ratio of dUTP:TTP used in the initial amplification reaction.

Figure 2 shows a gel showing the nucleic acid fragments produced when the method described herein was applied to exon 11 of BRCA2, eIF2, NS5 and p85nic.

Figure 3 shows the effect of UDG, APE and β-elimination treatment on NS5 PCR products comprising different levels of dUTP incorporation.

Figure 4 shows the PCR product produced after amplification of p85nic with 1% dUTP before and after fragmentation.

Figure 5 shows agarose gel electrophoresis of restriction digests of pCRBlunt/p85nic fragment clones and pCRT&-NT/p85nic fragment clones.

Figure 6 shows the analysis of the selectivity of the purification method for soluble vs insoluble protein. Samples of cell extract, Turbo-filtered cell extract And Ni-NTA eluate from purification trials of soluble cSti1, insoluble full length Gsk and the insoluble catalytic domain of Gsk were run on SDS-PAGE.

Detailed Description

Introduction

We have developed a method for identification of protein domains comprising two main steps: 1) production of a library of expression vectors that contain DNA fragments of defined size range that have been sampled essentially randomly from a particular target coding sequence; 2) screening of the library for clones that express soluble

protein domains. The first step employs an enzymatic fragmentation method based on the DNA base-excision repair pathway and the second step makes use of a protein purification method that is selective for soluble protein domains over insoluble protein fragments. The two key novel aspects of the methodology have been tested in two separate pilot feasibility studies: one involving the novel gene fragmentation aspects of the technology and another involving testing of the selectivity of the protein purification method for soluble proteins with tertiary structural integrity. These studies demonstrate that the DNA fragmentation method is efficient and reproducible, generating blunt-ended DNA fragments suitable for cloning. In addition, the fragment size range produced is found to be reproducible and solely a function of the ratio of dUTP:TTP used in the amplification of the PCR product. In a second aspect, these studies show the present protein purification method to be highly selective for soluble vs. insoluble protein and therefore suitable for screening of libraries of clones in order to identify those that produce soluble protein domains.

Materials and Methods

PCR

Initially four coding sequences were identified as potential targets for application of the "Domain hunting" method: human BRCA2 exon 11, yeast elongation initiation factor 2, Dengue virus type 1 NS5 and the N-SH2-Inter-SH2-C-SH2 region of the human signal transduction protein p85. Oligonucleotide primers were designed and synthesised for PCR amplification of each coding sequence. PCR was then performed using Taq DNA polymerase according to the manufacturers instructions

except that dGTP, dCTP, dATP were used at a concentration of 200 μM each, and TTP and dUTP were used at a concentration of 198 μM and 2 μM respectively. PCR was therefore performed in the presence of a ratio of 99% TTP to 1% dUTP allowing incorporation of dUTP at an average of ~1% at any particular thymidine nucleotide position in the sequences. Thirty cycles of PCR were performed for each template and an annealing temperature 5°C below the theoretical melting temperature was used for each reaction. The extension time used for each reaction was 60 seconds per kilobase of full-length product.

Fragmentation of PCR products

Digestion with UDG and APE enzymes:

The fragmentation protocol is summarised in Figure 1. The above NS5 and p85nic PCR products were treated with UDG (New England Biolabs. Inc.) and APE enzymes as below. Nth and NFO were over-expressed in *E. coli* and purified to homogeneity. Two different APE enzymes were assessed for their cleavage efficiency, NFO and Nth.

PCR products were purified by agarose gel electrophoresis and gel extraction according to the manufacturer's instructions (Qiagen Inc.) and then incubated with 1U of UDG per microgram of DNA and 2 μl of 2 μg/μl APE (either Nth or NFO) per microgram of DNA at 37°C for 60 mins. Spermine tetrahydorchloride (Calbiochem Inc.) was then added to 0.2mM final concentration before incubating at 37°C for 30 mins and then 70°C for 15 mins and 4°C 2 mins. The product was then purified (PCR purification kit, Qiagen Inc.) and the purified DNA eluted in 1 mM Tris.HCl pH8.0. The product was then incubated with 1 unit of S1-nuclease per microgram of DNA at 37°C for 60 mins. The product was then purified by 1% agarose gel

electrophoresis and a block of gel corresponding to DNA products of 300-600bp was excised and purified by gel extraction as above. The above product was then treated with shrimp alkaline phosphatase using one unit of enzyme per microgram of DNA at 37°C for one hour before adding the same quantity of fresh enzyme and incubating for a further hour. The reaction was then heated to 65°C for 15 minutes to totally inactivate the alkaline phosphatase. The product was then purified (PCR purification kit, Qiagen Inc.) and the purified DNA eluted in 1 mM Tris.HCl pH8.0. This DNA was then used for blunt-end cloning as described below. Alternatively, for TA cloning using the pCRT7-NT-TOPO vector (Invitrogen, Inc.) a final incubation with Taq DNA polymerase was performed to add single adenine nucleotide to the 3′ ends of the products. This was performed by incubating the product for 15 minutes at 72°C in the presence of a conventional PCR reaction mixture, well known to those skilled in the art, but without primers.

Cloning of the DNA fragments

~100ng of the above fragmented p85nic coding sequence was cloned using three different vectors (pCRBlunt, pCR4Blunt-TOPO and pCRT7-NT-TOPO) according to the manufacturer's protocol (Invitrogen Inc.). The transformation reactions were plated onto LB agar plates containing either ampicillin (pCRT7-NT-TOPO) or kanamycin (pCRBlunt, pCR4Blunt-TOPO) depending on the vector used for transformation.

Analysis of clones

Plasmid minipreps were performed (Qiagen inc.) for ~ 40 clones derived from pCRT7-NT-TOPO/p85 fragment transformations and for ~20 clones derived from

pCRBluntTOPO and ~20 clones derived from pCRBlunt. pCRT7-NT-TOPO/p85 fragment derived plasmids were digested with EcoR1 and BamHI (New England Biolabs Inc.) and analysed by 1% agarose gel electrophosesis.

Plasmid samples were DNA sequenced using the Cambridge University Biochemistry Dept. DNA sequencing service and results analysed using Vector NTI (Informax Inc.).

Selective purification of folded protein

50 ml cultures of *E. coli* BL21(DE3) cells expressing soluble C-terminal domain of Sti1 (cSti1) (REF), or insoluble Gsk3 (full length and catalytic domain) (REF) were pelleted and resuspended in 5 ml of lysis buffer (50 mM NaH_2PO_4, 300 mM NaCl, 1 mM imidazole pH 8.0). 1mg ml^{-1} lysozyme and 10 µg of Rnase A were added and the lysate incubated on ice for 30 min. 0.5ml of the lysate was then passed through a Qiagen TurboFilter (8 strip) as described by the manufacturer (Qiagen Inc.). 200 µl of the cleared lysates were added to 20 µl of Ni-NTA magnetic beads in 96 well microtitre plates. The plates were then shaken for 60 min, the beads washed twice in lysis buffer containing 10 mM imidazole, and bound protein eluted with 50 µl lysis buffer containing 300 mM imidazole. 20µl aliquots of the whole cell extract, the turbo-filtered extract and eluate from the beads was analysed by SDS-PAGE.

Results

DNA fragmentation trials

We have performed computer modelling experiments to predict the size of fragments that would be produced by the present DNA fragmention method for different levels

of dUTP incorporation. These predicted that 1% dUTP incorporation would produce a fragment size range with a distribution centering around 500bp. Four different coding sequences ranging in size from ~1-3.1kb were therefore amplified by PCR using Taq DNA polymerase in the presence of 1% dUTP demonstrating that PCR is highly efficient under these conditions (Figure 2).

We have then compared NS5 PCR products amplified using different ratios of TTP:dUTP (100:0, 99:1 and 90:10) by treatment with UDG and APE and β-elimination (Figure 3). This indicates that as expected the PCR products with no dUTP incorporated are unaffected by this treatment while 1% dUTP products show some slight evidence of fragmentation and 10% dUTP products show considerable evidence of fragmentation. These results are as expected since this treatment of 1% dUTP products with UDG and APE and β-elimination should introduce single stranded one nucleotide gaps in the dsDNA at ~500bp intervals on average. Similarly treatment of 10% dUTP products should produce gaps at intervals of around 50bp on average. On agarose gel electrophoresis therefore the 1% dUTP products would migrate in essentially the same way as uncut 100% TTP products since the 65°C 15 minute incubation step used for β-elimination would not be expected to cause significant melting of strands with 500bp overlaps between single-nucleotide gaps. The 10% dUTP product would however be expected to have melted significantly and then reannealed to produce a mixture of smaller annealed products consistent with that observed.

The whole fragmentation method (Figure 1) has been applied to 1% dUTP p85nic PCR product (Figure 2). This has been repeated using different APE enzymes and with

different lengths of incubation always yielding the same size distribution of product ranging from ~100bp to 1.2kb with maximum band intensity centred around 500bp as predicted (Figure 4). This process has been scaled up reproducibly for fragmentation of ~10 μg of DNA, indicating that generation of quantities of product sufficient for production of large libraries of clones according to the present invention is feasible.

Cloning

Transformation of *E. coli* cells with p85nic fragment cloning reactions was successful using three different cloning approaches: pCRBlunt ligation; pCR4Blunt-TOPO cloning; and pCRT7-NT-TOPO cloning. TOPO cloning of fragmented p85nic insert DNA into both pCR4Blunt-TOPO and pCRT7-NT-TOPO produced around 250 colonies per 100 ng of insert used. Blunt-end ligation of fragmented p85nic DNA to pCRBlunt produced ~1000 colonies at 16oC and 120 colonies at 37°C per 100 ng of fragmented DNA. These results indicate that a substantial proportion of the DNA fragments produced as described are blunt ended as expected. Cloning of DNA fragments produced by the method using the above cloning methods is therefore of sufficiently high efficiency to allow generation of libraries of thousands of clones.

Characterisation of cloned p85nic fragments

Restriction characterisation of plasmid DNA derived from clones generated by both TOPO cloning and blunt end ligation indicated that >90% of clones contained an insert and the distribution of the sizes of inserts correlated closely with the size range of p85nic DNA fragments used for cloning (Figure 5). DNA sequencing of the cloned DNA inserts suggests that the fragments appear

to be sampled in an essentially random manner from the p85nic coding sequence. No nucleotide substitutions have yet been detected by DNA sequencing, indicating that as expected the method is not inherently mutagenic. DNA sequencing of a large number of clones is necessary in order to accurately measure the randomness of sampling, frequency of mutation.

Selective purification of folded protein

In order to assess the selectivity of the purification method for folded protein versus unfolded or aggregated protein we have applied the purification method to a set of well-characterised proteins with known solubility properties. Cultures of *E. coli* BL21(DE3) cells expressing soluble C-terminal domain of Sti1 (cSti1), or insoluble Gsk3 (full length and catalytic domain) were harvested and the cells lysed enzymatically before passing through a Qiagen TurboFilter as described by the manufacturer (Figure 6). This step cleared the cell lysates and significantly reduced the amount of the insoluble Gsk in the lysate, but did not effect the level of the soluble cSti1. Further reduction of the quantity of insoluble constructs was seen following Ni-NTA magnetic bead purification. The cleared lysates were then purified using Ni-NTA magnetic beads in 96 well microtitre plates. The whole cell extract, the turbo-filtered extract and the Ni-NTA eluate were then analysed by SDS-PAGE showing that the recovery of the soluble cSti1 is at least 100 times more efficient than the insoluble constructs. The difference in the level of recovery of soluble vs. insoluble recombinant protein demonstrates that this purification method is highly selective for soluble folded protein over insoluble/misfolded protein over a wide dynamic range.

This purification approach will therefore allow sensitive detection of soluble folded protein fragments or domains over insoluble misfolded fragments and therefore allow identification of regions of protein sequence that correspond to folded protein.

Conclusions

The gene fragmentation study provided verification of incorporation of dUTP into the target gene by PCR, fragmentation of the target gene, robust control of the range of fragment sizes generated and efficicnet cloning of the fragments. We have tested the efficiency of PCR in the presence of dUTP for four different coding sequences. We have then compared the behaviour of PCR products prepared in the presence of different ratios of TTP:dUTP by treatment with uracil DNA glycosylase (UDG) and two different apurinic/apyrimidinic endonucleases (APE). This demonstrated that fragmentation occurs only to uracil containing PCR products and that the size of the fragments produced corresponds directly to the dUTP:TTP ratio used in the PCR amplification step. We selected the p85nic coding sequence for further analysis by the above enzymes and also for subsequent treatment with spermine and S1 nuclease. This demonstrated that fragments of p85nic of the size range predicted in theory for 1% dUTP incorporation were indeed produced. This also showed that as predicted these fragments were blunt ended since they could be cloned efficiently by blunt end cloning methods. A method for identification of soluble protein fragments or domains that can be efficiently expressed and purified from bacteria has been established and validated using several targets of well-characterised solubility properties. Coupling of the DNA fragmentation/cloning aspects with the soluble protein

domain identification aspects of the method therefore provides a holistic method for generation of vectors for high-level soluble expression of newly discovered protein domains. These vectors can then be used directly for production of large quantities of soluble protein domains for structural and functional studies, without the need for any subsequent genetic manipulation or optimisation of protein expression or purification.

References

The references cited herein are all expressly incorporated by reference.

Cohen, S. L. (1996) Structure 4 (9), 1013-1016.

Finucane, M. D., Tuna, M., Lees, J. H. and Woolfson (1999) *Biochemistry*, 38, 11604-11612.

Kawasaki, M. and Inagaki, F. (2001) *Biochem. Biophys. Res. Commun.* 280 (3), 842-844.

Moriki, T., Kuwabara, I., Liu, F. T. and Maruyama, I. N. (1999) *Biochem. Biophys. Res. Commun.* 265 (2), 291-296.

Sambrook J, Fritsch, EF, Maniatis, T (1989) *Molecular Cloning: A Laboratory Manual*, 2nd ed, pp 5.33-5.86.

Savva, R, McAuley-Hecht, K, Brown, T, Pearl, LH (1995) *Nature*, 373, 487-493.

Savva, R, Pearl, LH (1995) *Nature Structural Biology*, 2, 752-757.

Panayotou, G, Brown, T, Barlow, T, Pearl, LH, Savva, R (1998) *J Biol Chem*, 273, 45-50.

Barrett,TE, Savva, R, Panayotou, G, Barlow, T, Brown, T, Jiricny, J, Pearl, LH (1998) *Cell*, 92, 117-129.

Barrett, TE, Schärer,OD, Savva,R, Brown,T, Jiricny,J, Verdine,GL, Pearl,LH (1999) *EMBO J*, 18, 6599-6609.

Greagg, MA, Fogg, MJ, Panayotou, G, Evans, SJ, Connolly,

B, Pearl, LH (1999) *Proc Natl Acad Sci*, 96, 9045-9050.

Bailly V, Verly WG (1989) *Biochem. J.* 259, 761-768.

Wigley, W. C., Stidham, R. D., Smith, N. M., Hunt, J. F. and Thomas, P. J. (2001) *Nature Biotechnology* 19 (2) 131-136.

Claims:

1. A method of producing a library of nucleic acid fragments, the nucleic acid fragments encoding one or more portions of a polypeptide, and identifying fragments in the library encoding soluble protein domains, the method comprising:

amplifying a nucleic acid sequence encoding the polypeptide in the presence of a non-native nucleotide so that the non-native nucleotide is incorporated into the amplified product nucleic acid sequence at a frequency related to the relative amounts of the non-native nucleotide and its corresponding native nucleotide, if present;

contacting the product nucleic acid sequence with one or more reagents capable of recognising the presence of the non-native nucleotide and cleaving the product nucleic acid sequence or excising the non-native nucleotide, thereby producing nucleic acid sequences encoding fragments of the polypeptide;

expressing a library of the nucleic acid fragments to produce the protein domains encoded by the fragments, wherein the protein domains are expressed as fusions with an affinity tag; and

separating soluble proteins using the affinity tag.

2. A method for producing a library of nucleic acid fragments, the nucleic acid fragments encoding one or more portions of a polypeptide, the method comprising:

amplifying a nucleic acid sequence encoding the polypeptide in the presence of a non-native nucleotide so that the non-native nucleotide is incorporated into an amplified product nucleic acid sequence at a frequency related to the relative amounts of the non-native nucleotide and its corresponding native nucleotide, if

present;

contacting the product nucleic acid sequence with one or more reagents capable of recognising the presence of the non-native nucleotide and cleaving the product nucleic acid sequence or excising the non-native nucleotide, thereby producing nucleic acid sequences encoding fragments of the polypeptide.

3. The method of claim 1 or claim 2, wherein the step of amplifying the nucleic acid sequence is carried out using PCR using a non-native deoxynucleotide, either alone or in a mixture of the non-native and native nucleotide.

4. The method of claim 3, wherein the non-native nucleotide is uracil or 3-methyl adenine.

5. The method of any one of the preceding claims, wherein the nucleic acid sequence is present in a cDNA library, a RNA library or a sample of genomic DNA.

6. The method of any one of the preceding claims, wherein the nucleic acid fragments of the polypeptide are between 200 and 1200 nucleotides in length.

7. The method of any one of the preceding claims, wherein the nucleic acid sequence is sampled on average about every second nucleotide to produce the nucleic acid fragments.

8. The method of any one of the preceding claims, wherein the reagent capable of recognising and cleaving the product nucleic acid sequence at the non-native nucleotide is an enzyme which can recognise the presence

of the non native nucleotide and cleave the nucleic acid sequence at or around the modified nucleic acid sequence.

9. The method of claim 8, wherein the enzyme is a DNA glycosylase or an endonuclease.

10. The method of any one of the preceding claims, wherein the non-native nucleotide is deoxyuracil and the enzyme is apurinic/apyrimidinic endonuclease (APE), catalysed by uracil-DNA glycosylase (UDG).

11. The method of any one of the preceding claims, further comprising amplifying the nucleic acid fragments.

12. The method of any one of the preceding claims, wherein in the library of protein domains, the protein domains comprise a constant portion and a portion sampled by the amplifying and contacting steps.

13. The method of any one of the preceding claims, further comprising ligating the nucleic acid fragments into expression vector(s).

14. The method of claim 13, further comprising transforming host cells with the expression vectors to produce a library of host cells capable of expressing fragments of the polypeptide.

15. The method of claim 14, further comprising expressing the nucleic acid sequences encoding fragments of the polypeptide and optionally isolating the polypeptide fragments thus produced.

16. A method of identifying soluble protein domains, the method comprising:

expressing a library of nucleic acid fragments to produce the protein domains encoded by the fragments, wherein the protein domains are expressed as fusions with an affinity tag; and

separating soluble proteins using the affinity tag.

17. The method of any one of the preceding claims, wherein the polypeptide fragments are expressed to include a protease cleavage site.

18. The method of any one of the preceding claims, wherein the polypeptide fragments are expressed to include an affinity tag.

19. The method of claim 18, wherein the affinity tag is a peptide which is less than 15 amino acids in length.

20. The method of claim 18 or claim 19, wherein the affinity tag is fused to the C-terminus of the protein fragments.

21. The method of any one of claims 18 to 20, wherein the affinity tag is polyhistidine, a Flag or Glu epitope, a S-tag, calmodulin binding peptide or ribonuclease S.

22. The method of claim 21, wherein the affinity tag is a His_6-tag.

23. The method of any one of claims 14 to 22, further comprising releasing the soluble protein domains from the cells.

24. The method of claim 23, wherein the step of releasing the protein is carried out under non-denaturing conditions.

25. The method of claim 24, wherein the non-denaturing condition comprise the use of enzymes or non-denaturing detergents.

26. The method of any one of claims 14 to 25, further comprising filtering out unbroken cells, cell debris and insoluble material.

27. The method of any one of claims 14 to 26, further comprising clarifying the lysates by centrifugation.

28. The method of any one of claims 14 to 27, further comprising purifying cells transformed with different proteins in parallel by affinity chromatography.

29. The method of any one of the preceding claims, wherein the step of separating the soluble protein domains is carried out by contacting the library of protein domains with a solid phase having a binding partner of the affinity tag immobilised thereon.

30. The method of claim 29, wherein the binding partner is:

(a) metal ions such as Ni^{2+} or Co^{2+} for binding a polyhistidine affinity tag; or

(b) anti-Flag antibodies for binding a Flag or Glu epitope affinity tag; or

(c) streptavidin for binding a S-tag affinity tag; or

(d) calmodulin in the presence of Ca^{2+} for binding a

calmodulin binding peptide affinity tag; or

(e) aporibonuclease S for binding a ribonuclease S affinity tag.

31. The method of any one of the preceding claims, further comprising assaying for the presence of soluble protein fragments.

32. The method of claim 31, wherein the step of assaying is carried out using anti-tag ELISA, SDS-PAGE or LC-ESI-MS.

33. The method of claim 31 or claim 32, wherein the step of assaying for the soluble protein domains comprises quantifying the protein expression level of one or more or the protein domains.

34. The method of any one of the preceding claims, further comprising identifying or sequencing the soluble proteins.

35. The method of any one of the preceding claims, further comprising contacting the library of fragments or domains with:

(a) one or more candidate compounds to determine whether the candidate compound binds to and/or modulates an activity of a protein fragment or domain present in the library; and/or

(b) one or more test proteins to determine whether a protein fragment or domain present in the library binds to and/or modulates an activity of the test protein.

36. The method of any one of the preceding claims, further comprising contacting two or more libraries of

soluble protein fragments or domains to determine whether binding or modulation of activity occurs between the protein fragments or domains present in the libraries.

37. The method of claim 36, wherein the method is employed to determine which portions of the proteins used to construct the libraries are involved in binding and biological activity.

38. The method of any one of claims 35 to 37, wherein the method is used to determine binding between a ligand and a receptor, an enzyme and a substrate, an antibody and an antigen, or a small molecule and a protein.

39. A library produced by the method of any one of the preceding claims.

40. The library of claim 39, which is:

(a) a library of nucleic acid fragments of a parent nucleic acid sequence, wherein the nucleic acid fragments have a size range between 200 and 1200 nucleotides in length and are preferably sampled from the parent nucleic acid sequence on average about every second nucleotide; or

(b) a library of expression vectors which comprise a plurality of the nucleic acid fragments as set out in (a), wherein each fragment is ligated to a nucleic acid sequence encoding an affinity tag and optionally one or more further sequences to direct the expression of the nucleic acid fragment and the affinity tag; or,

(c) a library of host cells transformed with the expression vectors as defined in (b); or

(d) a library of polypeptide fragments produced by expressing the nucleic acid sequences set out in step

(a), wherein each polypeptide is coupled to an affinity tag.

5

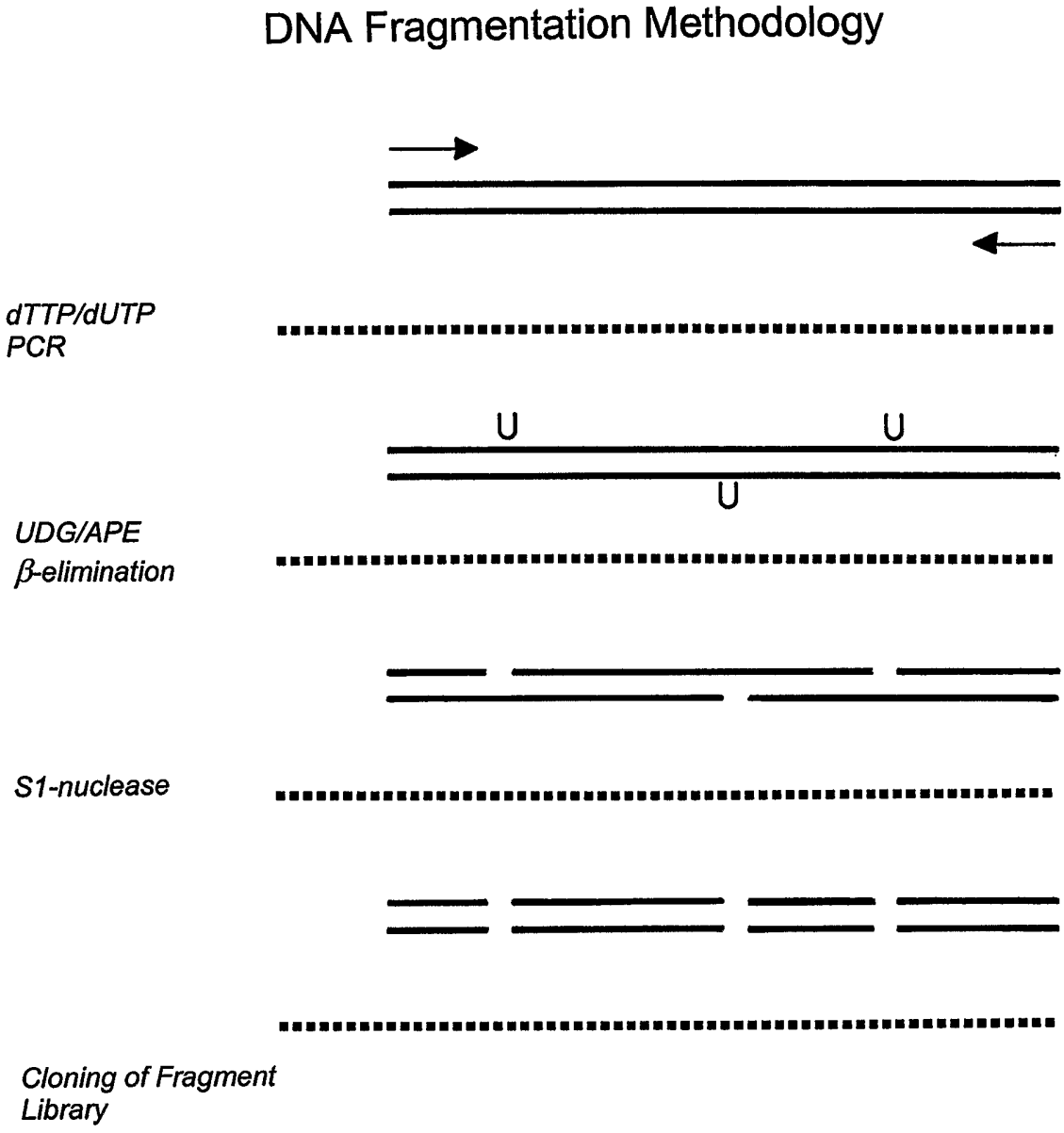
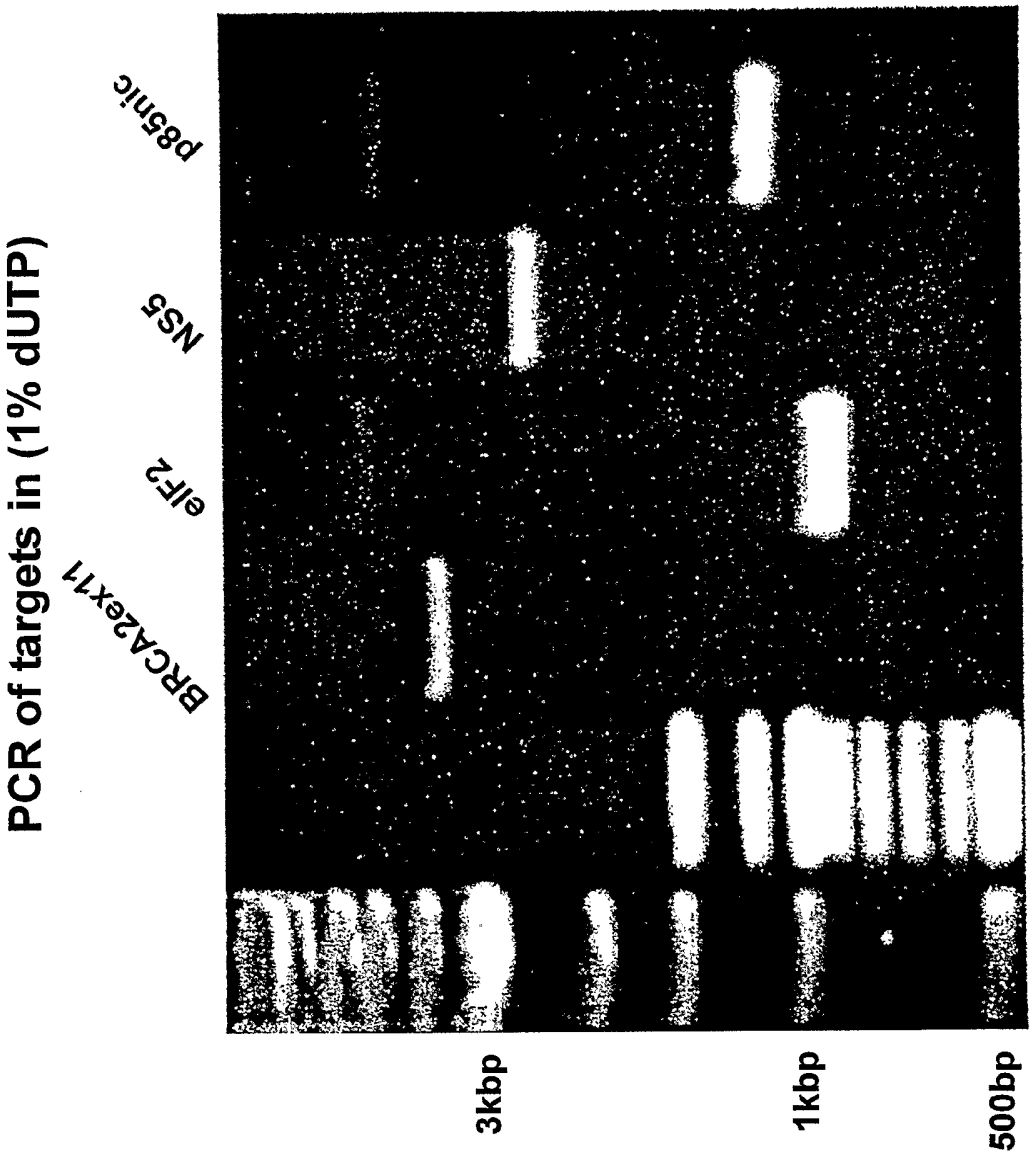
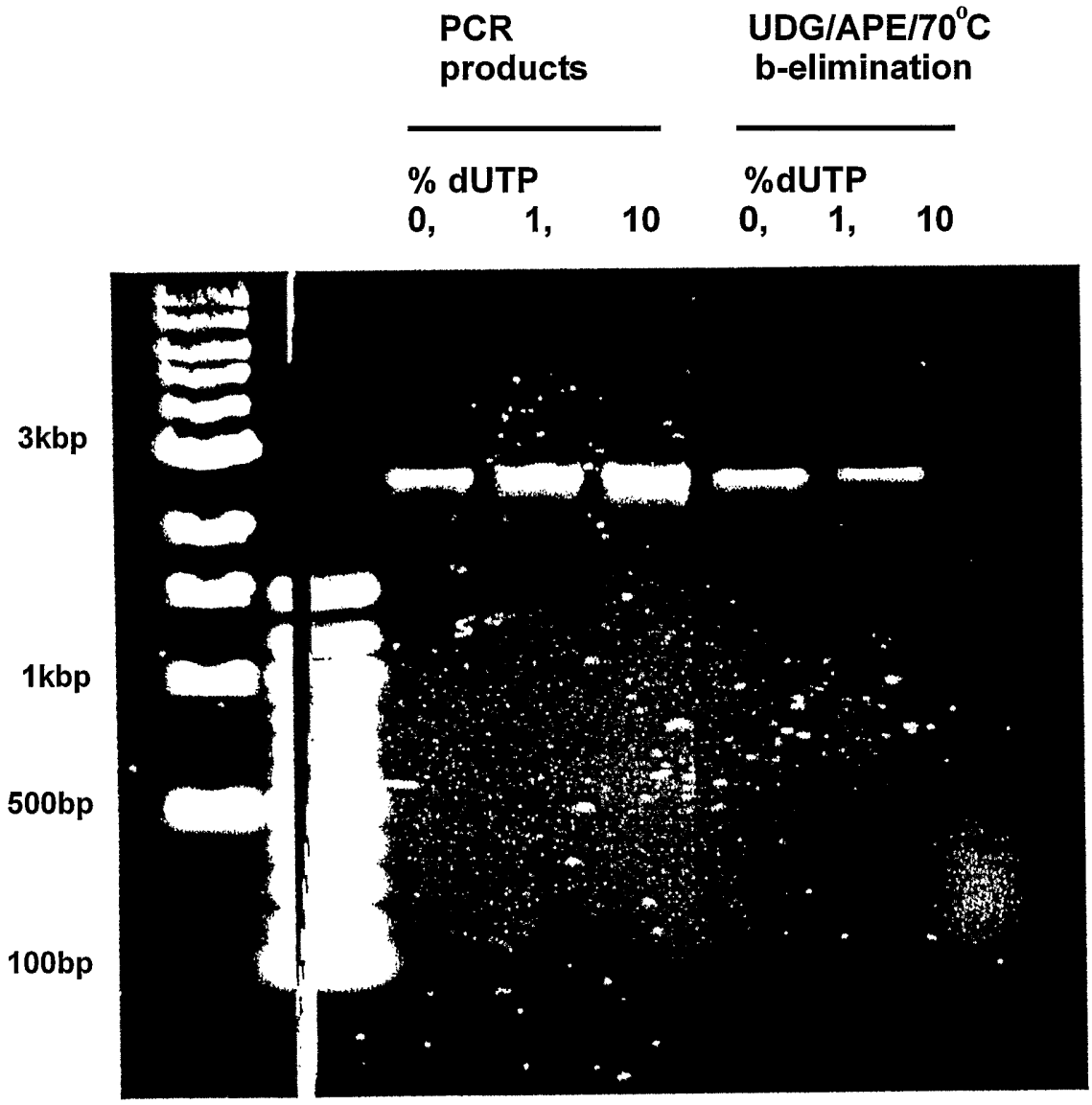

DNA Fragmentation Methodology
dTTP/dUTP
PCR
U
U
U
UDG/APE
β-elimination
S1-nuclease
Cloning of Fragment
Library

Fig. 1

Fig. 2

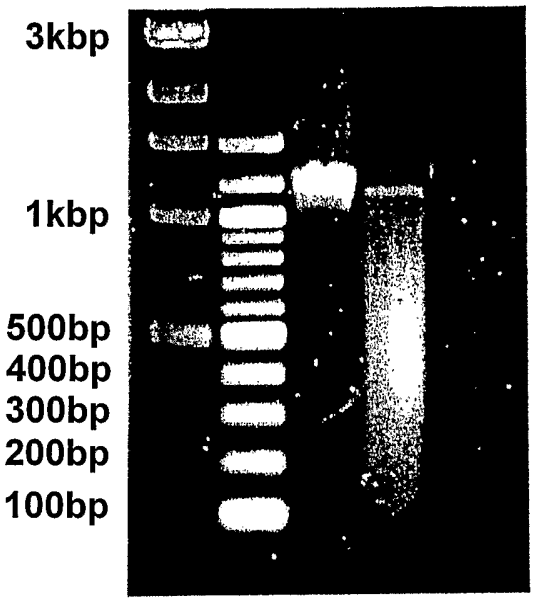


Lane 1, 1kb markers. Lane 2, 100bp markers. Lane 3, BRCA2 exon 11 PCR product (1% dUTP). Lane 4, eIF2 PCR product (1% dUTP). Lane 5, NS5 PCR product (1% dUTP). Lane 6, p85nic PCR product (1% dUTP).



Lane 1, kb markers. Lane 2, 100bp markers. Lanes 3-5, NS5 PCR products comprising 0,1 and 10% incorporation of dUTP. Lanes 6-8, NS5 PCR products comprising 0, 1 and 10% incorporation of dUTP after application of the UDG, APE and B-elimination steps of the method.

Fig. 3



Lane 1, 1 kb markers. Lane 2, 100bp markers. Lane 3,p85nic 1%dUTP PCR product. Lane 4, p85nic 1%dUTP PCR product after application of the fragmentation method

Fig. 4

Fig. 5

Lane 1, 1kb markers. Lane 2, 100bp markers. Lane 3-18,EcoRI digests of pCRBlunt-p85nic fragment clones. ECoR1 digestion releases the inserts from pCRBlunt therefore the bands between 300-1200 correspond to p85nic fragment inserts.

Lane 1, 1kb markers. Lane 2, 100bp markers. Lane 3-18,EcoR1/BamHI digests of pCRT7-NT-p85nic fragment clones. EcoR1/BamHI double digestion releases the inserts from pCRT7-NT therefore the bands between 300-1200 correspond to p85nic fragment inserts.

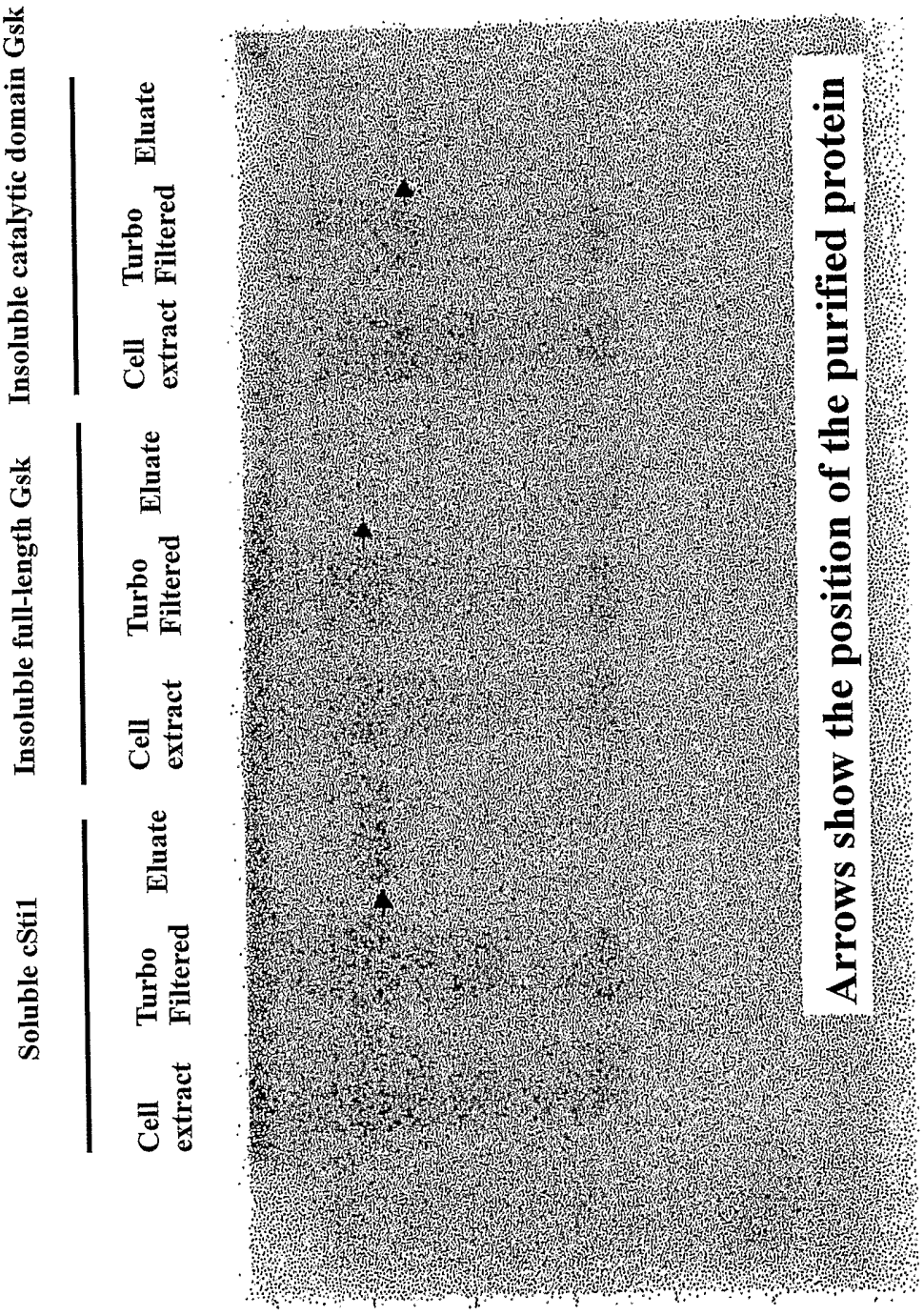

Soluble cStil
Cell extract
Turbo Filtered
Eluate
Insoluble full-length Gsk
Cell extract
Turbo Filtered
Eluate
Insoluble catalytic domain Gsk
Cell extract
Turbo Filtered
Eluate
Arrows show the position of the purified protein

Fig. 6