**PCT** WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

| (51) International Patent Classification 6 : C12Q 1/68 | A1 | (11) International Publication Number: **WO 98/15652** |
|---|---|---|
| | | (43) International Publication Date: 16 April 1998 (16.04.98) |

(21) International Application Number: PCT/GB97/02734

(22) International Filing Date: 6 October 1997 (06.10.97)

(30) Priority Data:
9620769.1 4 October 1996 (04.10.96) GB

(71) Applicant *(for all designated States except US)*: BRAX GENOMICS LIMITED [GB/GB]; 13 Station Road, Cambridge CB1 2JB (GB).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: SCHMIDT, Gunter [DE/GB]; Houghton Manor, Houghton, Cambs PE17 2BQ (GB). THOMPSON, Andrew, Hugin [GB/GB]; 25 Knoll Park, Alloway, Ayr KA7 4RH (GB).

(74) Agents: DANIELS, Jeffrey, Nicholas et al.; Page White & Farrer, 54 Doughty Street, London WC1N 2LS (GB).

(81) Designated States: AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CU, CZ, DE, DK, EE, ES, FI, GB, GE, GH, HU, IL, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, UA, UG, US, UZ, VN, YU, ZW, ARIPO patent (GH, KE, LS, MW, SD, SZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, 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: NUCLEIC ACID SEQUENCING BY ADAPTATOR LIGATION

(57) Abstract

A method for sequencing nucleic acid, which comprises: (a) obtaining a target nucleic acid population comprising nucleic acid fragments in which each fragment is present in a unique amount and bears at one end a sticky end sequence of predetermined length and unknown sequence, (b) protecting the other end of each fragment, and (c) sequencing each of the fragments by (i) contacting the fragments with an array of adaptor oligonucleotides under hybridisation conditions, each adaptor oligonucleotide bearing a label, a sequencing enzyme recognition site, and a known unique base sequence of same predetermined length as the sticky end sequence, the array containing all possible base sequences of that predetermined length; removing any unhybridised adaptor oligonucleotide and recording the quantity of any hybridised adaptor oligonucleotide by detection of the label, then repeating the cycle, until all of the adaptors in the array have been tested; (ii) contacting the hybridised adaptor oligonucleotides with a sequencing enzyme which binds to the recognition site and cuts the fragment to expose a new sticky end sequence which is contiguous with or overlaps the previous sticky end sequence; (iii) repeating steps (i) and (ii) for a sufficient number of times and determining the sequence of the fragment by comparing the quantities recorded for each sticky end sequence.

*FOR THE PURPOSES OF INFORMATION ONLY*

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **AL** | Albania | **ES** | Spain | **LS** | Lesotho | **SI** | Slovenia |
| **AM** | Armenia | **FI** | Finland | **LT** | Lithuania | **SK** | Slovakia |
| **AT** | Austria | **FR** | France | **LU** | Luxembourg | **SN** | Senegal |
| **AU** | Australia | **GA** | Gabon | **LV** | Latvia | **SZ** | Swaziland |
| **AZ** | Azerbaijan | **GB** | United Kingdom | **MC** | Monaco | **TD** | Chad |
| **BA** | Bosnia and Herzegovina | **GE** | Georgia | **MD** | Republic of Moldova | **TG** | Togo |
| **BB** | Barbados | **GH** | Ghana | **MG** | Madagascar | **TJ** | Tajikistan |
| **BE** | Belgium | **GN** | Guinea | **MK** | The former Yugoslav Republic of Macedonia | **TM** | Turkmenistan |
| **BF** | Burkina Faso | **GR** | Greece | | | **TR** | Turkey |
| **BG** | Bulgaria | **HU** | Hungary | **ML** | Mali | **TT** | Trinidad and Tobago |
| **BJ** | Benin | **IE** | Ireland | **MN** | Mongolia | **UA** | Ukraine |
| **BR** | Brazil | **IL** | Israel | **MR** | Mauritania | **UG** | Uganda |
| **BY** | Belarus | **IS** | Iceland | **MW** | Malawi | **US** | United States of America |
| **CA** | Canada | **IT** | Italy | **MX** | Mexico | **UZ** | Uzbekistan |
| **CF** | Central African Republic | **JP** | Japan | **NE** | Niger | **VN** | Viet Nam |
| **CG** | Congo | **KE** | Kenya | **NL** | Netherlands | **YU** | Yugoslavia |
| **CH** | Switzerland | **KG** | Kyrgyzstan | **NO** | Norway | **ZW** | Zimbabwe |
| **CI** | Côte d'Ivoire | **KP** | Democratic People's Republic of Korea | **NZ** | New Zealand | | |
| **CM** | Cameroon | | | **PL** | Poland | | |
| **CN** | China | **KR** | Republic of Korea | **PT** | Portugal | | |
| **CU** | Cuba | **KZ** | Kazakstan | **RO** | Romania | | |
| **CZ** | Czech Republic | **LC** | Saint Lucia | **RU** | Russian Federation | | |
| **DE** | Germany | **LI** | Liechtenstein | **SD** | Sudan | | |
| **DK** | Denmark | **LK** | Sri Lanka | **SE** | Sweden | | |
| **EE** | Estonia | **LR** | Liberia | **SG** | Singapore | | |

# NUCLEIC ACID SEQUENCING BY ADAPTATOR LIGATION

## Field of the Invention

The invention relates to a method for sequencing nucleic acid, especially DNA.

## Background to the Invention

Various methods for sequencing have been developed but most are slow and complicated to automate like Sanger's chain termination method or Maxam's chain degradation method. Others have technical difficulties remaining to be overcome like the single base sequencing methods to be discussed below. This is discussed in Brenner PCT/US95/12678 pg 1-2.

According to the method of chain termination sequencing (Sanger et al, Proc. Natl. Acad. Sci. USA 74, 5463 - 5467, 1977), modified nucleotides can be used to terminate polymerase extension of nucleic acid being copied from a template strand. To determine the sequence of a template strand four dideoxynucleotides are needed corresponding to the four normal bases. Template strands are added to a polymerisation medium containing all four normal nucleotides and one of the four dideoxynucleotides, which is labeled, usually with a fluorescent dye. The dideoxynucleotide in each medium is at a concentration such that it has a small but defined probability of being incorporated into an extending copy of the template rather than its corresponding normal nucleotide. This terminates chain extension for this fragment. If all the fragments in a particular medium are separated on a sequencing gel, which resolves nucleic acids to a difference in length of one nucleotide, fragments corresponding to termination at every occurrence of the base to which the dideoxynucleotide corresponds should be observed. If a gel for each base is run then there should be observed a band for each nucleotide in the template and the nucleotide sequence should be determined.

This method is limited to templates of about 1500 bp. It is automatable but there is a time bottleneck in the running of sequencing gels which is technically difficult to overcome. Gels

also have assorted problems of their own, such as band-broadening due to temperature effects, compressions due to secondary structure in the template nucleic acids and inhomogeneities in the separation gel.

Conceptually similar in principle to chain termination sequencing, the method of chain degradation sequencing (Maxam et al, Proc. Natl. Acad. Sci. USA **74**, 560 - 564, 1977) relies on chemical reagents which specifically cleave a DNA template at a specific base. Radiolabeled templates are cleaved in separate reactions with each reagent and separated on a sequencing gel. The pattern of bands is essentially the same as with chain termination but shifted by one base.

The cleavage reagents are, however, not totally specific, so this is a fairly 'noisy' system. It suffers from the same problems as the chain termination method too.

There are variations on methods of single-base sequencing, one example of which is described here. Others can be found in the references given below:

- Cheeseman, U.S. patent 5,302,509.

- Tsien et al, International Patent Application WO 91/06678.

- J.D. Harding and R.A. Keller, Trends in Biotechnology **10**, 55 - 58, 1992.

- Rosenthal et al, International Patent Application WO 93/21340.

- Canard et al, Gene **148**, 1 - 6, 1994.

- Metzker et al, Nucleic Acids Research **22**, 4259 - 4267, 1994.

The method of Harding and Keller acts on immobilised DNA templates. The templates are single stranded and constructed from

analogues of normal bases bearing unique fluorescent labels. The immobilised templates are cleaved with a 5′ to 3′ exonuclease to release bases into a flowing medium that is run through a fluorimeter that detects which base is present at the highest concentration in the medium. As long as the exonucleases cleave off bases simultaneously from each copy of a nucleic acid simultaneously, the signal at any time should correspond essentially to the last base cleaved. The sequence of bases flowing past the fluorimeter thus corresponds to the sequence of the fluorescent template. This sort of approach has a potential to be extremely rapid, limited only by the processivity of the exonuclease used, which could be of the order of 100 to a 1000 bases a second.

This method requires base analogs that are distinguishable by some means from each other, preferably by fluorescence. These analaogs must be incorporated into a template as normal bases and be cleaved by an exonuclease as normal bases. Alternatively polymerases and exonucleases must be engineered to recognise the analogs. Either way is a major technical obstacle. Furthermore, the requirement for simultaneity of cleavage of corresponding bases from a population of immobilised template allows only a small margin of error which will severely limit the length of sequence that can be determined by this approach. These technical obstructions remain to be overcome.

The alternative to ensuring that the exonuclease remains in step with numerous copies of a single template is to analyse only a single molecule at a time. This approach has severe technical difficulties such as manipulating single templates, developing fluorescent dyes with a large quantum yield and robustness to ensure every cleaved base is detected and developing corresponding detectors.

Sequencing by hybridisation to chips, grids and arrays is an approach described in J. Biomolecular Structure and Dynamics 9, 399 - 410, 1991. Arrays of single-stranded oligonucleotides can

be constructed representing for example, every possible combination of the 4 bases in an 8 bp oligonucleotide. Each point on an array would correspond to one such oligonucleotide. A single-stranded template with a fluorescent label can be hybridised to to such an array. Every overlapping linear sequence of 8 bp that is contained in the template will be represented on the array and the template should hybridise to every point corresponding to each 8 bp sequence that defines it. The fluorescence from every point on the array can then be determined and the sequence of the target reconstructed.

The problem with this sort of approach is the size of arrays required to sequence a nucleic acid template of reasonable length unambiguously. Arrays are expensive and technically difficult to construct.

Direct analysis of Sanger Ladders by mass spectrometry is an approach described by Köster *et al* in Nature Biotech 14, 1123-1128 (1996). This approach requires determination of the mass of each component of a Sanger ladder.

Analysis of Sanger sequence ladders directly by mass spectrometry has the potential to be extremely rapid but there is a severe problem with the 'read-length' that can be obtained using this approach. This is due to the fact that DNA is highly fragmentory in the mass spectrometer and is also poly-ionic, so each DNA species will be found in multiple ionisation states, which gives rise to highly complex spectra. Furthermore the mass resolution of most appropriate mass spectrometers does not permit sequencing of more than about 30 to 40 bases. This problem grows massively as the linear length of DNA sequence analysed increases.

The fragmentation problem and poly-ionisation problem are related - protonation of the bases in DNA is believed to induce fragmentation. Various DNA analogues exist which are less easily protonated, reducing the spectral complexity problem but the mass resolution limit is less trivial to overcome.

**Summary of the Invention**

The present invention provides a method for sequencing nucleic acid, which comprises:

(a) obtaining a target nucleic acid population comprising nucleic acid fragments in which each fragment is present in a unique amount and bears at one end a sticky end sequence of predetermined length and unknown sequence,

(b) protecting the other end of each fragment, and

(c) sequencing each of the fragments by

(i) contacting the fragments with an array of adaptor oligonucleotides under hybridisation conditions, each adaptor oligonucleotide bearing a label, a sequencing enzyme recognition site, and a known unique base sequence of same predetermined length as the sticky end sequence, the array containing all possible base sequences of that predetermined length; removing any unhybridised adaptor oligonucleotide and recording the quantity of any hybridised adaptor oligonucleotide by detection of the label, then repeating the cycle, until all of the adaptors in the array have been tested;

(ii) contacting the hybridised adaptor oligonucleotides with a sequencing enzyme which binds to the recognition site and cuts the fragment to expose a new sticky end sequence which is contiguous with or overlaps the previous sticky end sequence;

(iii) repeating steps (i) and (ii) for a sufficient number of times and determining the sequence of the fragment by comparing the quantities recorded for each sticky end sequence.

The label may be any label suitable for the purpose, such as a fluorescent label or a mass label. Each label may comprise a mass label associated with a corresponding known base sequence for identifying the corresponding base sequence in mass

spectrometry. In one embodiment, each adaptor oligonucleotide is labelled with an associated mass label which is uniquely resolvable in mass spectrometry from the other labelled adaptor oligonucleotides. According to this embodiment it is preferable that each adaptor oligonucleotide is composed of nucleotide analogues which are resistant to fragmentation in the mass spectrometer.

In another embodiment, each mass label is cleavably attached to its corresponding adaptor oligonucleotide and uniquely resolvable in mass spectrometry. The mass label may be attached to the adaptor oligonucleotide by a cleavable linker and may be cleaved under any appropriate cleavage conditions such as photocleavage conditions or chemical cleavage conditions.

The mass spectrometry may be effected using a mass spectrometer with orthogonal time of flight or array detector geometry.

According to one embodiment, the fragments are contacted in step (*i*) with the array of adaptor oligonucleotides in a cycle wherein the cycle comprises sequentially contacting each adaptor oligonucleotide of the array with the fragments.

Preferably, the target nucleic acid population is subjected to sorting into sub-populations according to their sticky end sequences and each of the sub-populations is subjected to steps (b) and (c).

In one embodiment, where the target nucleic acid is genomic DNA, each fragment may be produced by differential application.

Preferably, the predetermined length of this base sequence of the sticky ends is from 3 to 5,more preferably 4.

The sequencing enzyme preferably comprises a type IIs restriction endonuclease. The target nucleic acid population may comprise heterogeneous nucleic acid fragments._

The other end of each fragment may be protected by ligation with an immobilisation adaptor oligonucleotide.

**Brief description of the drawings**

FIGURE 1 shows a cloning vector for template sequences;
FIGURE 2 shows PCR amplification of template DNA;
FIGURE 3 shows immobilisation of amplified template DNA;
FIGURE 4 shows a method of differential amplification of template DNA fragments;
FIGURE 5 shows a method of producing protected DNA fragments with termini for sequencing;
FIGURE 6 shows the action of Fok1;
FIGURE 7 shows cutting behaviour of typical adapters according ot the invention;
FIGURE 8 shows an adapter cycle according to one embodiment of the invention; and
FIGURE 9 shows graphs of the effects of PEG and Ficol on ligation at ATTA and GCCG.

**DESCRIPTION OF THE PREFERRED EMBODIMENTS**

**Simultaneous sequencing of sorted populations of nucleic acids by adapters:**

This invention provides a method capable of simultaneously sequencing a heterogeneous population of nucleic acid fragments. The technology is compatible with numerous methods of template preparation. The invention however provides a novel preferred strategy for sequencing large DNA molecules that limits the need for biological sub-cloning hosts and vectors. In outline the sequencing process may be summarised:

1. Generation of a mixed nucleic acid population;
2. Sort molecules into subsets; and
3. Sequence molecules within subsets simultaneously.

The sequencing method described here allows one to produce nucleic acid fragment populations in a reproducible manner that

can then be sorted into subsets and finally sequenced by an oligonucleotide adapter based technique. The sequencing method described requires double stranded templates. The sequencing technique is most effective with immobilisation of the nucleic acid templates at one terminus, the other terminus must be accessible to adapters.

The sequencing steps use adapter molecules to generate and probe the sequence of terminal single-stranded overlaps of immobilised nucleic acid fragments. Single-stranded overlaps are generated in a cyclical process preferably through the use of type IIs restriction endonucleases. Recognition sites for these enzymes are provided by adapters at the terminus of a template sequence. The position of the recognition site is arranged so that digestion in the presence of the type IIs endonuclease exposes an ambiguous sticky end in the unknown sequence of the template. The resultant ambiguous sticky ends generated in template sequences are probed as heterogeneous sets and sequence information is determined by measuring the quantity of label detected from correctly hybridised adapters. The sequence of individual fragments is determined by comparing quantities of label for each adapter in each cycle of the sequencing process with quantities derived in previous and subsequent cycles. The invention provides a method for analysing heterogeneous sub-populations of nucleic acids without spatially resolving them. This is achieved by a signal acquisition and signal processing procedure that allows sequences to be identified on the basis of their relative quantities.

This process does not require traditional gel methods to acquire sequence information. Since the entire process takes place in solution and is an iterative process, the steps involved could be performed by a liquid-handling robot or a microfluidics system.

*Type IIs Restriction Endonucleases:*

Type IIs restriction endonucleases have the property that they recognise and bind to a specific sequence within a target DNA molecule, but they cut at a defined distance away from that sequence generating single-stranded sticky-ends of known length but unknown sequence at the cleavage termini of the restriction products.

For example, the enzyme **fok1**, generates an ambiguous sticky-end of 4 bp, 9 bp downstream of its recognition sequence. This ambiguous sticky-end could thus be one of 256 possible 4-mers. (see figure 6). Numerous other type IIs restriction endonucleases exist.

**Sequencing large nucleic acid molecules:**

It is not necessary to sequence an entire molecule at once to determine its sequence, which is fortunate as it is a practical impossibility, at the moment, to sequence molecules as large as chromosomes. It is calculated that any given sequence 17 bp long should be unique within the human genome. Similar calculations can be performed for genomes that are of different sizes. This consideration means that large nucleic acids or entire genomes can be sequenced by degradation into short overlapping fragments, > 17 bp in length, which can then be sequenced and the total genome sequence can thence be reconstructed using software to determine contig overlaps.

**Preparing a large Nucleic Acid for Sequencing:**

To sequence a complete nucleic acid of significant size is practically very difficult. A preferred method for use with this invention requires fragmentation of the target nucleic acid followed by molecular sorting into sub-populations that are small enough to allow simultaneous sequencing. This preferred method requires the use of adapters. Two adapter based sorting methods are described here. One requires the use of a type IIs restriction endonuclease or a similar system for generating

ambiguous sticky-ends in double stranded DNA while the second uses a primer and DNA amplification based approach to sorting.

Establishing a single sequencing terminus for generic nucleic acids is a requirement that faces users of this technology. The approach taken will clearly be determined by what is known about the nucleic acid to be sequenced. An adapter based approach suitable for an unknown nucleic acid is described here.

*Fragmentation of large nucleic acids:*

Nucleic acids may be fragmented in numerous ways which may be either directed or random. For the purposes of sequencing large nucleic acids, an approach which generates numerous fragments that overlap randomly is favoured in conventional sequencing strategies for large nucleic acid templates, due to its redundancy and relative simplicity. Obviously this sort of approach requires multiple copies of the target nucleic acid to ensure that all sequences are represented in the population of fragments with unambiguous overlaps with other fragments.

Random fragmentation will work excellently with certain embodiments of this invention but to try and reduce redundant sequencing more controlled fragmentation of a target nucleic acid could be used. One might fragment the target nucleic acid with a relatively high stringency type II restriction endonuclease or a type IIs restriction endonuclease. A set of relatively high stringency restriction endonucleases can be used to generate sets of overlapping fragments. In this way one would hope to generate overlapping contigs in a more economical manner than random fragmentation.

Random fragmentation can be achieved with mild digestion of the target nucleic acid by DNAseI or sonication. Generally, blunt-ended fragments are generated by this approach.

**Automated preparation of heterogeneous template populations with**

**known template frequencies:**

In order to produce a high throughput DNA sequencing technology the automation of the production of the sequencing template is essential. The following is an outline which describes an automated method of producing sequencing templates using conventional techniques.

For a large scale sequencing project, for example a whole bacterial genome or a full YAC clone, the DNA is first sub-cloned into a library. The process of producing a library of this sort can be done in-house or by commercially available services, such as that provided by Clonetech. The DNA is fragmented (e.g. by restriction enzyme digestion or sonication) to sizes in range of a few hundred bases and then sub-cloned into a cloning vector of choice. Because each fragment in the library is flanked by the same vector sequence a standard set of flanking PCR primers can be used to PCR amplify each fragment. Using the same PCR primers for each fragment also helps to normalise the efficiency of each PCR reaction as primer sequence is one of the most important factors affecting amplification efficiency. (see figure 1)

The library is then transfixed into an appropriate bacterial strain and the bacteria plated out onto selective agar plates. Individual colonies (each containing an unique fragment contained within the cloning vector) are then picked by a colony picking robot (which are commercially available). Each picked colony is then spiked into a unique PCR reaction, set up on a microtitre plate for example, and each fragment is PCR amplified using the standard primer set which flank the insert. One of the primers used in this reaction must be biotinylated which will allow the subsequent capture of the amplified fragment. (see figure 2)

Following the PCR amplification, a known amount of each of the amplified fragments can be captured on a streptavidin coated surface by its biotinylated primer. By controlling the amount of available streptavidin a specific amount of PCR product can be

captured. (This does, however, rely on all the primers being incorporated into the amplification products. This should only require a simple primer titration optimisation experiment as PCR reactions using clones are highly efficient.)

Different protocols can be used for this purpose, for example streptavidin coated magnetic beads or streptavidin coated wells of a microtitre plate. When using beads, which will bind 1 pmol of biotin per ul of beads, adding 5 ul of beads and the appropriate buffer to the PCR reaction will capture 5 pmol of the amplified fragment. The use of beads also allows the capture of different quantities of individual amplified fragments. By adding differing amounts of beads to separate amplification reactions prior to pooling them, one can, for example, create a heterogeneous population with 1pmol of fragment 1, 4 pmol of fragment 2, 10 pmol of fragment 3 and so on. Alternatively streptavidin coated wells of a microtitre plate could also be used by transferring each amplification reaction to a unique well of the microtitre plate. Commercially available streptavidin coated plates usually have a maximum binding capacity of between 5 to 20 pmol of biotin. Therefore, the amount of amplified fragment captured in each well is determined by the binding capacity of that plate.

Following capture, excess amplified fragments are then washed away, the double stranded PCR product is denatured with either alkali or heat (or both) (to free the non-biotinylated strand). The non-biotinylated strand is then washed away and this leaves a single stranded template immobilised in the well or tube ready to be used in a sequencing reaction. (see figure 3)

If simultaneous sequencing of heterogeneous templates is to be performed quantification of template must be stringent. This can be achieved by labelling one of the primers. After the amplification reaction has been performed, the biotinylated fragments can be captured on an avidinated substrate and washed. The number of copies of template present can be determined by

measuring the retained fluorescence. Appropriate dilution of the amplified template can be performed if desired before sequencing. This gives an additional level of control over and above the capture steps.

The vector sequences at either terminus of the DNA can be designed to bear distinct primer sequences. This would ensure that one terminus can be readily identified as a sequencing terminus and one terminus could be designated as the immobilisation terminus. A unique termination sequence at the immobilisation terminus would identify when the clone had been sequenced. Type IIs restriction endonuclease sites in the immobilisation terminus sequence can additionally permit molecular sorting. The sequencing terminus can be engineered to carry a recognition sequence for the type IIs restriction endonuclease to be used as the sequencing endonuclease. These vector sequences can additionally provide primer sequences to permit amplification of template and amplification based sorting.

**Adapter based techniques for template preparation:**

Rather than using cloning vectors to provide sequencing and immobilisation termini it is possible to use an adapter based approach that is more amenable to automation. One can ligate adapters bearing primer binding sites to the nucleic acid fragments generated by a variety of fragmentation processes.

The precise method of ligation will depend on how the fragments are generated. If fragments of the target nucleic acid are generated by using a type IIs restriction endonuclease, adapters with sticky-ends complementary to subsets of the possible sticky-ends that would be generated by the fragmentation endonuclease, can be ligated to the resultant fragments. These adapters could carry designed primer sites that would allow much greater control of the amplification step. The combinations of subsets of sticky-ends of the primer adapters will determine which subsets of fragments are amplified and how large these subsets of fragments

are. This will allow much greater control over the PCR amplification steps. See Figures 4 and 5.

To maximise the differences in frequency between fragments within an amplified set, one need only alter the quantities of primers corresponding to the primer binding sites present in each adapter added to the PCR incubator. The combination of adapters at the termini of the nucleic acid fragments should increase the variation in frequency of fragments, exacerbating the known inhomogeneities in PCR amplification.

*PCR of genomic DNA:*

The use of controlled fragmentation and amplification, outlined above, could conceivably allow specific amplification of DNA subsets directly from genomic DNA. This potentially offers a novel strategy for genomic sequencing that avoids cloning into biological hosts and vectors. If one could reliably amplify genomic subsets of fragments then the biological steps of present sequencing strategies could be avoided with large savings in time and resources, and also eliminating the unreliability of using biological vectors. A set of primers would also be a rather more manageable way to access genomic fragments: primers would not need to be physically maintained due to the ease of synthesising short oligonucleotides whereas clones must be carefully cultured to ensure their availability and continued integrity.

The parallel sequencing process described here lends itself to this sort of cloning strategy because of its ability to simultaneously sequence heterogeneous populations without spatial resolution of nucleic acids which conventional sequencing strategies cannot achieve. This means if a set of primers generated more than one fragment, the ability to sequence multiple templates simultaneously would allow one to determine the sequence without having to separate the amplified fragments.

*Ensuring that only the desired type IIs restriction endonuclease*

*sites in the target nucleic acid are available for sequencing:*

It is important to ensure no 'sequencing enzyme' binding sites are accessible or present in the template nucleic acid fragments prior to addition of adapters bearing the 'sequencing enzyme' binding site to the terminus of the molecule from which sequencing is to occur. Certain type IIs restriction endonucleases are sensitive to the methylation state of their recognition regions so to prevent unwanted sites being used by the sequencing endonuclease the target nucleic acid can be methylated prior to ligation of adapters bearing the sequencing endonuclease recognition site. Methylation can be achieved in the preparation of templates by use of 5-methyl cytosine in any amplification reactions. Use of unmethylated adapters would allow recognition sequences present in these to function but not those in the template.

An alternative, but less preferable, way to avoid this problem is to remove enzymatically the recognition sequence of the sequencing endonuclease from within the target nucleic acid population. The nucleic acid can be fragmented initially with the sequencing enzyme. This will generate 3 classes of fragments, one class with the sequencing enzyme recognition site at one terminus only, one class with the sequencing enzyme recognition site at both termini and a third class with the site at neither termini.

A complete adapter set, i.e. corresponding to all sticky-ends, can be added to the restriction fragment population. The adapter would bear the recognition site for the sequencing enzyme. Addition of the sequencing enzyme to a population of fragments with these adapters can have two results. If a given terminus has a recognition site already then the sequencing enzyme can cleave either at the adapter site or at the more internal site. There is a 50 % chance of either cleavage event occurring. At other sites where there is no internal site, clearly, terminal bases must be lost by this process. Since with each round of this process only half of the internal sites will be removed, the

process must be repeated at least 7 times to ensure removal of the sequencing enzyme recognition sequence from at least 99 % of the fragments in the population. Thus fragment size may be significantly reduced if a sequencing enzyme is used that cleaves at a significant distance from its recognition site.

*Establishing distinct termini in a population of nucleic acids:*

An important facet of this technology is immobilisation of nucleic acids at one terminus. This requires that a randomly generated fragment have directionality, i.e. it requires two distinguishable termini. This can be achieved using adapters. Two types of adapters are required to identify two distinct termini. Adapters for a simple protocol are shown below:

Adapter 1

Biotin —— **NNN GGATG NNNNNGATC**
**NNN CCTAC NNNNN**

Adapter 2

**NNN GGCC NNNNGATC**
**NNN CCGG NNNN**

The pair of adapters shown can be ligated to a fragment generated by Sau3AI. The first adapter bears a recognition sequence for fok1 while the second adapter bears a restriction site for BsuRI. BsuRI is methylation sensitive and generates blunt-ended fragments. If one synthesises template DNA with S-methyl cytosine but uses adapters with ordinary DNA, only the adapter will be cleaved by this will leave fragments amenable to blunt end ligation. Adapter 1 provides immobilisation and the recognition site for the sequencing endonuclease. A simple protocol for

generating distinct termini would be as follows:

- The first step is fragmentation of a large number of copies of a large nucleic acid, preferably with an ordinary type II restriction endonuclease to generate known sticky ends, such as Sau3AI.

- The resultant fragments can then be ligated to adapters. If the fragments are treated with ligase in the presence of the two types of adapters above, this will generate fragments of three types: fragments with both ends carrying adapter 1, fragments with both ends carrying adapter 2 and thirdly fragments carrying adapter 1 at one end and adapter 2 at the other. Statistically the third type of fragment will be in the majority.

- If the immobilisation effector on adapter 1 is biotin then the fragments carrying adapter 1 can be immobilised on a solid phase matrix derivitised with avidin. The fragments carrying adapter 2 at both ends can be washed away and those fragments carrying two immobilisation adapters will also be immobilised.

- Cleavage with the type II restriction endonuclease whose binding site is carried by adapter 2 will allow blunt end ligation of a new adapter to one terminus of the fragments bearing both types of adapter.

- At this stage the immobilised fragments can be removed from the solid phase matrix. Biotin/streptavidin interactions can be disrupted by acid.

- Fragments that bore both adapters can be captured by the new terminus generated by cleavage of adapter 2. Capture requires a further adapter which can be immobilised allowing fragments with adapter 1 at both termini to be washed away. Alternatively the 'capture' adapter can introduce a primer sequence. Adapter 1 can additionally provide a known primer sequence to permit the captured fragments to be differentially amplified.

A more complex protocol which allows molecular sorting could be achieved using adapter 2 to provide a second type IIs recognition site. (BspMI in the example below)

Adapter 2

**NNNGCAGGT NNNN**
**NNNCGTCCA NNNNGATC**

If the cleavage step after immobilisation is performed with this would generate fragments with ambiguous sticky ends at one terminus which can again be captured by adapters complementary to the sticky ends generated but one can select at this stage which sticky ends to capture. An entire array of all possible adapters can be generated to allow all fragments to be captured and isolated. A hybridisation array on a glass surface would allow spatial sorting. An alternative method would use the adapter sequence to perform differential amplification.

*Additional sorting:*

Once a fragment population has been amplified and distinct termini established for each fragment, an arbitrary degree of sorting can be performed. The 'capture' adapter used above can provide another terminal type IIs restriction endonuclease site. This will allow another set of ambiguous sticky-ends to be generated allowing further sub-sorting until the nucleic acid fragment population is of the correct size for unambiguous sequence determination.

This sorting process above generates, for a 4 bp ambiguous sticky-end, 256 sub-populations. This may be generate nucleic acid populations small enough to begin sequencing or further sub-sorting may be necessary.

In order to begin to sequence the sorted nucleic acids, they must be treated with the sequencing enzyme to expose a new ambiguous

sticky-end at the sequencing terminus.

*Parallel Sequencing of Subsets of Nucleic Acid Fragments with Adapters:*

*Sequencing a single molecule by ligation of adapters:*

The actual sequencing method is essentially sequencing by hybridisation and can be understood first by explaining it for the case of a single nucleic acid. Consider a single nucleic acid, immobilised at one terminus to a solid phase substrate, and which has an adapter at the other terminus bearing the recognition site for the type IIs restriction endonuclease chosen for sequencing. Digestion in the presence of fok1 will generate a 4 bp ambiguous sticky-end.

To determine the sequence of that sticky-end one can probe the immobilised nucleic acid with an adapter molecule. This would be an oligonucleotide carrying a sticky-end with one, known, sequence of 4 bp of the possible 256. The adapter would additionally carry a label, e.g. a fluorescent tag, and a binding site for the desired type IIs restriction endonuclease to be used to sequence the immobilised nucleic acid. If the adapter is complementary to the ambiguous end of the target nucleic acid, it will hybridise and it will then be possible to ligate the adapter to the target. The immobilised matrix can then be washed to remove any unbound adapter. To determine whether the adapter has been ligated to the immobilised target, one need only measure the fluorescence of the matrix. This will also reveal how much of the adapter has hybridised, hence the amount of immobilised DNA. Other means of detecting hybridisation might be used in this invention. Radio-labelled adapters could be used as an alternative to a fluorescent probe, so also could dyes, stable isotopes, tagging oligonucleotides, enzymes, carbohydrates, biotin amongst others. If the adapter is not complementary to the ambiguous sticky-end of the target nucleic acid and only one label is available, then a second adapter can be tried and the

above process repeated until all 256 possible adapters have been tested. (This process is shown in the 'adapter cycle' - see figure 8).

Clearly one of the adapters will have to be complementary to the ambiguous end. Once this has been found, then the terminus of the target nucleic acid will carry also a binding site for the sequencing endonuclease that will allow cleavage of the target nucleic acid exposing further bases for analysis and the above process can be repeated for the next 4 bp of the target. This iterative process can be repeated until the entire target nucleic acid has been sequenced.

A more effective method of labelling appropriate for use with this invention is 'mass labelling'. Cleavable mass labels are described in patent GB9700746.2. This describes methods for generation and use of labels that are readily detectable in a mass spectrometer. Mass labelling permits the generation of large numbers of labels. This would permit 256 labels to be generated allowing all 256 probes for a 4 base pair overlap to be tested simultaneously rather than sequentially. This has advantages in a hybridisation based sequencing method as a competitive binding system avoids some of the problems of different binding energies of different 4 base sequences. GB 9700746.2 describes tagging of nucleic acid probes with cleavable mass labels. These labels may be cleaved from the probe at various stages in a probing assay, but a preferred method of cleavage is during the ionisation process. For use with the sequencing method described here various methods are possible. After the exposed sticky ends of a template are probed with labelled adapters one is left, after washing away non-ligated adapters, with a template terminated with a labelled adapter. One can use labels that are photocleavable, thermo-labile or acid labile, for example, which can be removed at this stage and analysed in a mass spectrometer. Alternatively one can cleave the adapter from the template with the appropriate type IIs restriction endonuclease whose recognition site is provided by the adapter. The cleaved adapter

can be analysed in a mass spectrometer and the mass label can be cleaved during ionisation.

Non-cleavable mass labels are also appropriate for analysis of the cleaved adapter. One needs only use sufficient labels to resolve adapters with the same mass in the mass spectrum.

For the purposes of sequencing, with 4bp ambiguous sticky-ends, one would need 256 adaptors each tagged via a non-cleavable linker, to give the whole adaptor, or preferably just the strand of the adapter without the probe sequence, a unique mass in the mass spectrum. After probing a template with an adaptor population the successfully ligated adapter can then be identified after washing away any unligated probe. To identify the ligated adapter, it is then digested from the template by the sequencing endonuclease. The released adaptor can then be analysed by electrospray mass spectrometry. Preferably, only the mass tagged strand should be analysed. A short liquid chromatography step with a denaturing solvent would allow the tagged strand to be separated from the untagged strand. HPLC or capillary electrophoresis separations would be appropriate. Such a separation would probably not be necessary, though. Denaturing the cleaved adaptor might be quite desirable, however. After cleavage, both strands of the adaptor will be extended by 4 bp. The probe strand will be extended by 4 unknown bases. The non-probe strand will be extended by 4 bases complementary to the probe overlap of the probe strand, and so will have a known mass, hence is the preferred strand for mass labelling. Certain 4mers have the same composition, GGCC, GCCG, GCGC, etc and need to be resolved as these will all give the same peaks in a mass spectrum. One need only add sufficient mass to resolve these uniquely. One can furthermore choose mass tags which should reveal mis-hybridisations as an incorrect base will be cleaved off the template and will be present in the extended sequence of the tagged probe. If probes appear shifted to points in the mass spectrum in which probes are not normally found one will detect mis-hybridisation. If the tags are chosen carefully it should

be possible to determine the identity of the mis-hybridised base. This should allow correction of hybridisation errors and correct assignment of sequence. Quantitation of probe should be improved if this source of error is removed. Type IIs restriction endonucleases are also known to sometimes cleave at incorrect positions. Such cleavages should also be identifiable with this approach as an extra base or one base too few will give a shifted mass spectral peak. This should again allow improvement in quantitation.

Analysis of small nucleic acid molecules within the mass spectrometer is very much less affected by fragmentation than the analysis of larger DNA molecule making this a highly appropriate method of analysis. If fragmentation were a concern one could synthesise adapters and template DNA with analogues that are resistant to fragmentation. N-7 deaza analogues of guanine and adenine are appropriate.

*Sequencing a Population of Nucleic Acid Fragments:*

The same process can be applied to a heterogeneous population of immobilised nucleic acids allowing them to be analysed in parallel. To be successful when applied to a population of nucleic acids, this method relies on the fact that statistically 1 out of 256 molecules within the total population will carry each of the possible 4 bp sticky-ends at any particular site after cleavage with a sequencing enzyme such as **fokI**. If one sub-sorts the template population into manageable subsets of less than 256 fragments, one would expect that almost all will have different ambiguous sticky-ends (there is about a 1 in 1000 chance of there being 2 distinct cDNAs having the same initial 4 bp sticky end) so for most purposes one can assume that a hybridisation signal corresponds to a single cDNA type.

The positioning of the recognition site for the sequencing endonuclease in the adapter will determine whether the next 4 bp exposed are the next 4 bp in the sequence. Or they may overlap

partially with the last four base pairs thus giving partially redundant information or they may be further downstream missing out a few bases, thus only sampling the sequence of the immobilised target nucleic acid.(See figure below, sequential bases can be exposed with adapter 1 while bases are sampled at intervals by adapter 2. With adapter 3 redundant information is acquired. Adapter nucleic acid is shown in bold while fok1 binding sites are underlined). Whatever spacing is used, the spatial information relating the 4 bp oligonucleotides is retained. For the purposes of this invention redundant sequence data is desirable from the template nucleic acid in order to relate sequence information from each round of sequencing to the last round. This gives a small amount of redundancy, hence adapter 3 in figure 7 below is a preferred adapter construct.

Reconstructing Sequences of Target Nucleic Acids:

Repetitions of the adapter cycle will generate a matrix of quantities of label corresponding to each adapter corresponding to each adapter tested.

| Adapter | Cycle 1 | Cycle 2 | Cycle 3 | Cycle 4 |
|---|---|---|---|---|
| AAAA | 5 | 24 | 13 | 7 |
| AAAC | 10 | 5 | 9 | 13 |
| AAAG | 13 | 9 | 15 | 17 |
| ..... | .... | .... | .... | .... |
| TTTG | 7 | 13 | 17 | 10 |
| TTTT | 17 | 10 | 7 | 14 |

To reconstruct the sequences to which these quantities of label correspond, this invention also envisions algorithms for analysing such a data matrix. The algorithm attempts to identify a sequence on the basis of its frequency, i.e. a sequence present at a given frequency will have every sub-sequence present at the same frequency. The algorithm searches through each column of the matrix and attempts to resolve label quantities, that may be sums of sequence frequencies into atomic quantities such that the same set of atomic quantities appear in all columns. The algorithm

achieves this by comparing label quantities in a given column with those in the all the other columns. A given atomic quantity that appears in all columns is then assumed to correspond to a unique sequence.

If two sequences have the same n-mer at a particular point in the sequence, these can be resolved by the quantitative nature of this system in that the quantity of a particular n-mer in a particular ligation will be the sum of the quantities of the two sequences that share the n-mer at the same point. These can be largely resolved by comparison of one cycle with previous and subsequent ligation cycles to identify such sums. This is made particularly simple if the sequences that are being analysed have been amplified by PCR such that the sequence in the lowest quantity is present at not less than half the quantity of the sequence with the greatest frequency, that is to say if the frequency range of sequences lies between some quantity N and 2N. This means that any sum of frequencies will be greater than 2N and hence readily detectable.

If there is a known overlap between sequence samples, in embodiments that use adapters that generate overlapping sequence samples one has a certain amount of redundancy with which to account for errors. A one base overlap in samples will a quarter of the sequences in each column of the matrix with the next and previous columns.

**Implementation of the process:**

Practical details of implementing the process are described below.

**Adaptors, PCR Primers and Oligonucleotides:**

*Construction of Adaptors, Primers, etc:*

Details and reviews on the construction of oligonucleotides are available in numerous up to date texts, which should allow one skilled in the art to construct primers, adaptors and any other oligonucleotides required by the invention:

o Gait, M.J. editor, 'Oligonucleotide Synthesis: A Practical Approach', IRL Press, Oxford, 1990

o Eckstein, editor, 'Oligonucleotides and Analogues: A Practical Approach', IRL Press, Oxford, 1991

o Kricka, editor, 'Nonisotropic DNA Probe Techniques', Academic Press, San Diego, 1992

o Haugland, 'Handbook of Fluorescent Probes and Research Chemicals', Molecular Probes, Inc., Eugene, 1992

o Keller and Manack, 'DNA Probes, 2nd Edition', Stockton Press, New York, 1993

o Kessler, editor, 'Nonradioactive Labeling and Detection of Biomolecules', Springer-Verlag, Berlin, 1992.

**Conditions for Using Oligonucleotide Constructs:**

Details on effects of hybridisation conditions for nucleic acid probes can be found in be found in references below:

o Wetmur, Critical Reviews in Biochemistry and Molecular Biology, **26**, 227-259, 1991

o Sambrook et al, 'Molecular Cloning: A Laboratory Manual, 2nd Edition', Cold Spring Harbour Laboratory, New York, 1989

oPlease consider the attached claims and proposed response and let us know whether or not you are happy with them. If we do not hear from you by the **start of our business day 6th October 1997**, we shall take it that you are happy with the claims and response and file them at the European Patent Office without further revision.

Hames, B.D., Higgins, S.J., 'Nucleic Acid Hybridisation: A Practical Approach', IRL Press, Oxford, 1988

*Ligation:*

Ligation of adaptors is another critical aspect of the invention that must be considered. Chemical methods of ligation are known:

o Ferris et al, Nucleosides and Nucleotides **8**, 407 - 414, 1989

o Shabarova et al, Nucleic Acids Research **19**, 4247 - 4251, 1991

Preferably enzymatic ligation would be used and preferred ligases would be T4 DNA ligase, T7 DNA ligase, E. coli DNA ligase, Taq ligase, Pfu ligase and Tth ligase. References to the literature are given below:

o Lehman, Science **186**, 790 - 797, 1974

o Engler et al, 'DNA Ligases', pg 3 - 30 in Boyer, editor, 'The Enzymes, Vol 15B', Academic Press, New York, 1982

Protocols for use of ligases can be found in:

o Sambrook et al, cited above

o Barany, PCR Methods and Applications, **1**: 5 - 16, 1991

o Marsh et al, Strategies **5**, 73 - 76, 1992

*Phosphorylation of Nucleic Acids:*

When ligases and restriction endonucleases are used, there are changes made to the 5' phosphates of nucleic acid backbone sugar molecules. It may be advantageous to alter the phosphorylation state of adaptors or target nucleic acids in versions of the process. For example dephosphorylating the terminal 5' bases of immobilised cDNAs left after **fokI** cleavage to prevent cross-ligation of immobilised nucleic acids with complementary termini. Hence included are references to literature regarding use of phosphatases, kinases and chemical methods:

o Horn and Urdea, Tetrahedron Lett. **27**, 4705, 1986

o Sambrook et al, cited above

**Restriction Endonucleases:**

Numerous type IIs restriction endonucleases exist and could be used as sequencing enzymes for this process. Table 1 below gives a list of examples but is by no means comprehensive. A literary review of restriction endonucleases can be found in Roberts, R.,

J. Nucl. Acids Res. **18**, 2351 - 2365, 1988. New enzymes are discovered at an increasing rate and more up to date listings are recorded in specialist databases such as REBase which is readily accessible on the internet using software packages such as Netscape or Mosaic and is found at the World Wide Web address: **http://www.neb.com/rebase/**. REBase lists all restriction enzymes as they are discovered and is updated regularly, moreover it lists recognition sequences, isoschizomers of each enzyme, manufacturers and suppliers and references to them in scientific literature. The protocol would be much the same irrespective of the type IIs restriction endonuclease used but the spacing of recognition sites for a given enzyme within an adaptor would be tailored according to requirements and the enzymes cutting behaviour. (see figure n above)

| **Enzyme Name** | **Recognition sequence** | **Cutting site** |
|---|---|---|
| Fok1 | GGATG | 9/13 |
| BstFs1 | GGATG | 2/0 |
| SfaNI | GCATC | 5/9 |
| HgaI | GACGC | 5/10 |
| BbvI | GCAGC | 8/12 |

Table 1: A sample of type IIs restriction endonucleases

The requirement of the process is the generation of ambiguous sticky-ends at the termini of the nucleic acids being analysed. This could also be achieved by controlled use of 5′ to 3′ exonucleases. Clearly any method that achieves the creation of such sticky-ends will suffice for the process.

Similarly ordinary type II restriction endonucleases required by this invention can be found in the reference sources listed above. Details on methylation sensitivity and other means of controlling enzyme action can be found in the references given in REBase or can be acquired from the manufacturers.

Other means, however, of cleaving the immobilised nucleic acid might also suffice for this invention. Site specific chemical cleavage has been reported in Chu, B.C.F. and Orgel, L.E., Proc.

Natl. Acad. Sci. USA, 1985, 963 - 967. Use of a non-specific nuclease to generate blunt ended fragments might also be used. Preferably, though, a type II restriction endonuclease would be used, chosen for accuracy of recognition of its site, maximal processivity and cheap and ready availability.

**Solid Phase Supports:**

A full discussion of solid phase supports can be found in Brenner PCT/US95/12678 pg 12 - 14. This is an important issue in the use of fluorimetry to determine sequence abundance in that the design of supports will affect the acquisition of fluorescent signals which must be maximised for this process to be effective.

**Fluorimetry:**

Preferred embodiments of the process will use adaptors labeled with fluorescent labels. Detection of fluorescent signals can be performed using optical equipment that is readily available. Fluorescent labels usually have optimum frequencies for excitation and then fluoresce at specific wavelengths in returning from an excited state to a ground state. Excitation can be performed with lasers at specific frequencies and fluorescence detected using collections lenses, beam splitters and signal distribution optics. These direct fluorescent signals to photomultiplier systems which convert optical signals to electronic signals which can be interpreted using appropriate electronics systems.

Brenner PCT/US95/12678 pg 26 - 28 gives a full discussion.

**Liquid Handling Robotics:**

For this process to be practically useful, automation is essential and liquid handling robots can be acquired from various sources such as Applied Biosystems.

**Quantification and mass spectrometry:**

For the most part biochemical and molecular biological assays are quantitative. The mass spectrometer is not a simple device for quantification but use of appropriate instrumentation can lead

to great sensitivity. It must always be remembered that the ion count is not a direct measure of the source molecule quantity, the relationship is a complex function of the molecule's ionisation behaviour. Quantitation is effected by scanning the mass spectrum and counting ions at each mass/charge ratio scanned. The count is integrated to give the total count at each point in the spectrum over a given time. These counts can be related back to the original qunatities of source molecules in a sample. Methods for relating the ion count or current back to the quantity of source molecule vary. External standards are one approach in which the behaviour of the sample molecules is determined prior to measurement of unknown sample. A calibration curve for each sample molecule can be determined by measuring the ion current for serial dilutions of a sample molecule when fed into the instrument configuration being used.

Internal standards are probably the more favoured approach rather than external standards, since an internal standard is subjected to the same experimental conditions as the sample so any experimental vagaries will affect both internal control and sample molecule. To determine the quantity of a sample molecule, an internal standard of a known quantity is added to the sample. The internal standard is chosen to have a similar ionisation behaviour as the molecule being measured. Thus the ratio of sample ion count to standard ion count can be used to determine the quantity of sample as the ratio of qunatities should be the same. Choosing appropriate standards is the main difficulty with this approach. One must find a molecule that is similar but not identical in its mass spectrum. A favourable approach is to synthesise the sample molecule with appropriate isotopes to give a slightly different mass spectrum, for a molecule with the same chemical behaviour. This approach might be less desirable than external standards for use with large numbers of mass labels due to the added expense of finding or synthesising appropriate internal standards but will give better qunatification than external standards. An alternative to isotope labelling is to identify a molecule that has similar but not identical chemical behaviour as the sample in the mass spectrometer. Finding such

analogues is difficult and is a significant task for large families of mass labels.

A compromise approach might be appropriate though, since large families of mass labels will ideally be synthesised combinatorially, and will thus be related chemically. A small number of internal controls might be used, where each individual control determines the quantities of a number of mass labels. The precise relationship between internal standard and each mass label might be determined in external calibration experiments to compensate for any differences between them.

The configuration of the instrument is critical to determining the actual ion count itself, particularly the ionisation method and the separation method used. Certain separation methods act as mass filters like the quadrupole which only permits ions with a particular mass charge ratio to pass through at one time. This means that a considerable proportion of sample never reaches the detector. Furthermore most mass spectrometers only detect one part of the mass spectrum at a time. Given that a large proportion of the mass spectrum may be empty or irrelevant but is usually scanned anyway, this means a further large proportion of the sample is wasted. These factors may be a problem in detecting very low abundance ions but these problems can in large part be overcome by correct configuration of the instrumentation.

To ensure better quantification one could attempt to ensure all ions that are generated are detected. Mattauch-Herzog geometry sector instruments permit this but have a number of limitations. Sector instruments are organised into distinct regions, 'sectors', that perform certain functions. In general the ionisation chamber feeds into a free sector which feeds into an 'electric sector'. The electric sector esentially 'focusses' the ion beam which is divergent after leaving the ion source. The electronic sector also ensures the ion stream has the same energy. This step results in the loss of a certain amount of sample. This focussed ion beam then passes through a second free area into a magnetic sector which splits the beam on the basis of its mass charge ratio. The magnetic sector behaves almost like

a prism. A photographic plate can be placed in front of the split beam to measure the intensities of the spectrum at all positions. Unfortunately there is a limit on the dynamic range of these sorts of detector and it is messy and cumbersome. Better dynamic range is achievable with electron multiplier arrays, but at a cost of loss in resolution which is limited by how close together the elements of the array can be constructed. With a family of well characterised mass labels one would probably monitor only sufficient peaks to sample all the mass labels unambiguously. In general array detectors would allow one to simultaneously and continuously monitor a number of regions of the mass spectrum simultaneously, which might be applicable for use with well characterised mass label families. The limit on the resolution of closely spaced regions of the spectrum might restrict the number of labels one might use, though, if array detectors are chosen. For 'selected ion monitoring' (SIM) the quadropole has an advantage over many configurations in that the fields that filter ions can be changed with extreme rapidity allowing a very high sampling rate over a small number of peaks of interest.

Orthogonal TOF mass spectrometry:

An approach that is preferable to array geometries is the orthogonal time of flight mass spectrometer. This geometry that allows for very fast sampling of an ion stream followed by almost instantaneous detection of all ion species. The ion current leaving the source, probably an electrospray source for many biological applications, passes an electrode plate perpendicular to the current. This plate is essentially an electrical gate and is used to generate a repulsive potential which deflects the ion current 'orthogonally' into a time of flight mass analyser that uses a reflectron. The reflectron is essentially a series of circular electrodes that generate an increasingly repulsive electromagnetic fieldthat normalises ion energies and reflects the ion stream into a detector. The reflectron is a simple device that greatly increases the resolution of TOF analysers. Ions leaving the ion source will have different energies, faster ions will penetrate the repulsive field further than ions with a lower

energy and so will be delayed slightly with respect to the lower energy ions but since they will arrive slightly before the lower energy ions they will enter the TOF at roughly the same time so all the ions of a given mass charge ratio will arrive at the detector at roughly the same time. When the electrical gate is 'closed' to deflect ions into the TOF analyser, the timer is triggered. The flight time of the deflected ions is recorded and this is sufficient to determine their mass/charge ratio. The gate generally only sends a short pulse of ions into the TOF analyser at any one time. Since the arrival of all ions is recorded and since the TOF separation is extremely fast, the entire mass spectrum is measured effectively simultaneously. Furthermore, the gate electrode can sample the ion stream at extremely high frequencies so very little sample is required. For these reasons this geometry is extremely sensitve, to the order of a few femtomoles.

**Example 1:**

Ligation of adapters and cycles of digestion have been demonstrated - PCT/GB95/00109. The purpose of the following examples is to show that the ligation of labelled adaptors can give a quantitative signal that is proportional to the quantity of template present.

Three different PCR products are used to represent 3 different templates at different frequencies. The PCR product used for this are exons 14, 16 and 19 of the anion exchanger (AE1) as these PCRs have already been optimised in our laboratory. These are referred to as AE14, AE16 and AE19.

The products are captured to Dynalbeads (by incorporating a biotin in one of the PCR primers) and effectively represent captured cDNA. AE16 will be at half the concentration of AE14 and AE19 will be at one fifth the concentration of AE14.

*AE14 sequence*

ccaaagctgggagagaacagaatgccttggttttctgctgcagatcttccaggaccacccactacagaagac

ttataactacaacgtgttgatggtgcccaaacctcagggcccccctgcccaacacagccctcctctcccttgt

gctcatggccggtaccttcttctttgccatgatgctgcgcaagttcaagaacagctcctatttccctggcaa

gtcagcataccctcctcgcctgtccttgccaacactgc

*AE16 sequence*

ctgggagaatgccagggaaaggtctctgcctcccaccctcccaggcccagcccccaccctgtctctcacgtg

gtgatctgagactccaggaatatgaggatgaagaccagcagagcaggcagggcggaggcaaaatcatccaga

tgggaaactcggaacgcaagcccagtgggtggatgacccagccccgggctgaggagttgacaccttgaagcc

atcaggcaccgagagtttctgtgggaggggggtagcaggtaagaatgccaagggc

*AE19 sequence*

gtgataggcactgaccccagcctccgcctgcaggtgaagacctggcgcatgcacttattcacgggcatccag

atcatctgcctggcagtgctgtgggtggtgaagtccacgccggcctccctggccctgcccttcgtcctcatc

ctcactgtgccgctgcggcgcgtcctgctgccgctcatcttcaggaacgtggagcttcagtgtgtgagtggc

tgcctgggcctggggcacaagagctgggagcatgcg

Following capture, they are first digested with the frequent cutter Sau 3A1. This enzyme recognises the sequence GATC.

This provides the following 4bp overhangs of each of the products.

*AE14*

TTCCAGGACCACC...
CTAGAAGGTCCTGGTGG...

*AE16*

TGAGACTCCAGGAATAT...
CTAGACTCTGAGGTCCTTATA...

*AE19*

ATCTGCCTGGCAG...
CTAGTAGACGGACCGTC...

The following adaptor complimentary to the 4bp overhang revealed by Sau 3A1, and containing a Fok I site, is then ligated to the captured fragments.

*Adaptor SauFAM*

FAM - CTAGAGGACGATCGA.GGATG.
GATCTCCTGCTAGCT.CCTAC.CTAG

|

Fok I site

This produces the following sequences

*AE14*

FAM - CTAGAGGACGATCGA.GGATG.GATC.TTCCAGGACCACC...

GATCTCCTGCTAGCT.CCTAC.CTAG.AAGGTCCTGGTGG...

*AE16*

FAM - CTAGAGGACGATCGA.GGATG.GATC.TGAGACTCCAGGAATAT...

GATCTCCTGCTAGCT.CCTAC.CTAG.ACTCTGAGGTCCTTATA...

*AE19*

FAM - CTAGAGGACGATCGA.GGATG.GATC.ATCTGCCTGGCAG...
GATCTCCTGCTAGCT.CCTAC.CTAG.TAGACGGACCGTC...

These sequences are then digested with Fok I, which cuts at 9 and 13 bases from GGATG, and the following fragments are released into solution.

*AE14*

FAM - CTAGAGGACGATCGA.GGATG.GATC.TTCCA
GATCTCCTGCTAGCT.CCTAC.CTAG.AAGGTCCTG

*AE16*

FAM - CTAGAGGACGATCGA.GGATG.GATC.TGAGA

GATCTCCTGCTAGCT.CCTAC.CTAG.ACTCTGAGG

*AE19*

FAM - CTAGAGGACGATCGA.GGATG.GATC.ATCTG

GATCTCCTGCTAGCT.CCTAC.CTAG.TAGACGGAC

The cleaved fragments are then captured, through ligation, to 3 different wells of a microtitreplate each containing a specific adaptor simulating the first cycle of a sequencing reaction, providing the first 4 bases.

See below for full sequences

*For AE14 (adaptor Bbv14)*

```
Biotin-N-GCAGC.AGA
       N-CGTCG.TCT.CAGG
         |
         Bbv I site
```

*For AR16 (adaptor Bbv16)*

```
Biotin-N-GCAGC.AGA
       N-CGTCG.TCT.CCTC
```

*For AE19 (adaptor Bbv19)*

```
Biotin-N-GCAGC.AGA
       N-CGTCG.TCT.GTCC
```

Where N is a number of bases

This produces the following sequences:

*For AE14*

```
Biotin-N-GCAGC.AGA.GTCCTGGAAGATC.CATCC.AGCTAGCAGGAGATC
       N-CGTCG.TCT.CAGGACCTTCTAG.GTAGG.TCGATCGTCCTCTAG -FAM
```

*For AR16*

```
Biotin-N-GCAGC.AGA.GGAGTCTCAGATC.CATCC.AGCTAGCAGGAGATC
       N-CGTCG.TCT.CCTCAGAGTCTAG.GTAGG.TCGATCGTCCTCTAG -FAM
```

*For AE19*

Biotin-N-GCAGC.AGA.CAGGCAGATGATC.CATCC.AGCTAGCAGGAGATC

N-CGTCG.TCT.GTCCGTCTACTAG.GTAGG.TCGATCGTCCTCTAG-FAM

At this point the concentration can be measured through fluorescence of the FAM label and the first 4 bases (XXXX) determined.Successful ligation, measured by fluorescence therefore provides concentration information and the first 4 bases of each fragment.

**Adaptor Sequences and Preparation:**

*SauFam*

5'-FAM-CTAGAGGACGATCGAGGATG-3'

3'-GATCTCCTGCTAGCTCCTACCTAG-PO4-5'

*'Bbv' Adaptors*

Bbv14

5'BIOTIN-6C-CCTAGACTAGAGGACCGATCGAATCAGCAGCAGA-3'

3'-GATCTGATCTCCTGGCTAGCTTAGTCGTCCTCTCAGG-PO4-5'

Bbv16

5'BIOTIN-6C-CCTAGACTAGAGGACCGATCGAATCAGCAGCAGA-3'

3'-GATCTGATCTCCTGGCTAGCTTAGTCGTCCTCTCCTC-PO4-5'

Bbv19

5'BIOTIN-6C-CCTAGACTAGAGGACCGATCGAATCAGCAGCAGA-3'

3'-GATCTGATCTCCTGGCTAGCTTAGTCGTCCTCTGTCC-PO4-5'

*Cycling Aadptors*

C14

5'FAM-CAACTGTCCAGGATC-3'
3'-GTTGACAGGTCCTAGAAGG-PO4-5'

C16

5'FAM-CAACTGTCCAGGATC-3'
3'-GTTGACAGGTCCTAGACTC-PO4-5'

C19

5'FAM-CAACTGTCCAGGATC-3'
3'-GTTGACAGGTCCTAGTAGA-PO4-5'

BioFAMFok

5'BIOTIN-GGTCACTTAGATCGATCCATGAGGATGCTTCATTCTGATTCAGTCC-3'
3'-CCAGTGAATCTAGCTAGGTACTCCTACGAAGTAAGACTAAGTCAGG-FAM

BioG

5'BIOTIN-GCATCTGGAGTCTACAGTCGTCTATTGACG-3'
3'-CGTAGACCTCAGATGTCAGCAGATAACTGCCGGC-PO4-5'

GCCG

5'FAM-GCATCAGGATGTACAG-3'
3'-CGTAGTCCTACATGTCGCCA-PO4-5'

In the above adaptors the abbreviations used are as follows:

FAM- fluorescein
PO4 - phosphate

All primers were purchased from Oswell DNA Services.

All adaptors were made by heating 200ul of TE containing each primer at 20pmol/ul concentration at 90°C, in a Techne Dryblock and allowing the block to cool to room temperature over 2 hours. The adaptors were then incubated on ice for 1 hour and then frozen at -20°C until used.

Binding Bbv14,16, and 19 Adaptors to Microtitre plate

In order to capture the Fok 1 cleaved fragments to the 'Bbv' adaptors via ligation the 'Bbv" adaptors were bound to black, streptavidin coated 96 well microtitre plates (Boehringer Mannheim). This was achieved by incubating 10pmol of the appropriate adaptor in 35ul of 1xTE+0.1M NaCl in each well overnight at 4°C. Following the overnight incubation each well was washed 3 times with 50ul of 1xTE+0.1M NaCl. The 1xTE+0.1M NaCl was removed and 50ul of 1xligase buffer was added to each well and the plate was stored at 4°C untill used.

Plate capacity

To determine the binding capacity of each well 10pmol of BioFAMFok adaptor was bound to 8 wells by incubating 10pmol of the adaptor in 25ul of 1xTE+0.1M NaCl in each well overnight at 4°C. Following the overnight incubation each well was washed 3 times with 50ul of 1xTE+0.1M NaCl. A dilution of BioFAMFok (5, 2.5, 1.25, 0.675, 0.3375pmol) diluted in 1xTE+0.1M NaCl was added to a series of well and the fluorescence of the plate read in a Biolumin Microtiter plate Reader (Molecular Dynamics)

The following readings (expressed as Relative Fluorescent Units)

were obtained.

*Dilution wells*

5 pmol 74575 RFU

2.5pmol 35429 RFU

1.25pmol 16232 RFU

0.625pmol 9388 RFU

0.3375pmol 4807 RFU

*Wells incubated with 10pmol of adaptor and washed*

20872 RFU

21516 RFU

22519 RFU

21679 RFU

22658 RFU

21517 RFU

21742 RFU

22417 RFU

mean=21865

From these figures one can calculate that 21856 RFUs is equal to 1.5 pmol of BioFAMFok. These data agree with the capacity of the wells to bind biotinylated double stranded DNA (5pmol hybridised in 200ul) provided by Boehringer Mannheim technical help line.

Effect of Tween 20 on Ligation

The addition of 0.1% Tween 20 to the reaction buffer used with Fok 1 is claimed to reduce the exonuclease activity associated with this enzyme (Fok 1 data sheet - New England Biolabs). The following experiment was performed in order to determine if the addition of Tween would have any effect on the subsequent ligation of the cleaved fragements.

Nine reactions were set up with each set of three reactions each containing either 0, 0.05 or 0.1% tween in 25ul of 1xligase buffer, 10pmol BioG adaptor, 10pmol GCCG adaptor and 200ul ligase (New England Biolabs). One set of three reactions was set up as the above with 0.1%tween and no ligase. These were then incubated at 16°C for 1 hour and then each reaction transfered to a well of a black streptavidin coated microtitre plate (Boehringer Mannheim). The plate was incubated at room temperature for one hour and each well washed 3 times with 100ul of TES and the fluoresence measured in a Biolumin Microtite plate Reader (Molecular Dynamics).

The following readings (expressed as Relative Fluoresent Units) was obtained.

| Sample | 0% tween 20 | 0.05% tween 20 | 0.1% tween 20 | 0.1% tween 20 (no ligase) |
|---|---|---|---|---|
| 1 | 8592 | 8742 | 10213 | 3660 |
| 2 | 8083 | 8712 | 10605 | 3967 |
| 3 | 8720 | 8519 | 11598 | 3468 |
| means | 8465 | 8657.7 | 10805 | 3698 |

The above data demonstrate that the inclusion of 0.1% tween 20 increases ligation effieciency and therefore should not be detrimental to the ligation of the Fok 1 cleaved fragments to the 'Bbv" adaptors.

## PCR primers and Conditions and Purification

The 3 PCR products used to represent sequence templates at different concentrations were exons 14,16 and 19 from the human

erythrocyte anion exchanger gene located on chromosome 17q21-22. Primer sequences use to amplify exons 14,16 and 19

Exon 14
Forward primer
5'-GTATTTTCCAGCCCAAGCCAAAGCTGG-3'

Reverse primer

5'BIOTIN-GCAGTGTTGGCAAGGACAGGC-3'

Exon 16

Forward primer
5'BIOTIN-GCCCTTGGCATTCTTACCTGC-3'

Reverse primer

5'-CTGGGAGAATGCCAGGGAAAGG-3'

Exon 19

Forward primer

5'-GTGATAGGCACTGACCCCAG-3'

Reverse primer

5'BIOTIN-CGCATGCTCCCAGCTCTTGTGC-3'

The inclusion of biotin into one of the primers in each set will allow their capture to streptavidin coated beads (Dynal UK).

All PCR reactions were performed in 50ul containing 1xAmplitaq buffer (Perkin Elmer), 30pmol of forward and reverse primer,

200uM dNTPs, 1.25 units of Amplitaq (Perkin Elmer) and 100ng of human genomic DNA. The reactions were overlaid with 50ul of mineral oil and cycled on a Techne 'Genie' PCR machine with the following conditions.

Exon 14

1 cycle 95°C for 2 min
35 cycles 57.5°C for 45 sec

72°C for 1 min

95°C for 35 sec

1 cycle 72°C for 5 min

*Exon 16*

1 cycle 95°C for 2 min

35 cycles 52°C for 45 sec

72°C for 1 min

95°C for 35 sec

1 cycle 72°C for 5 min

*Exon 19*

1 cycle 95°C for 2 min

35 cycles 57.5°C for 45 sec

72°C for 1 min

95°C for 35 sec

1 cycle 72°C for 5 min

*Purification*

Excess primers and salts need to be removed before the PCR products are bound to DynaBeads, this is performed as described below.

10 reactions of each were pooled following PCR, separately, prior to purification. The PCR products were then ethanol precipitated by added 2.5 volumes of 100% ethanol and one tenth of a volume of 3M sodium acetate. The solution was then incubated at -20°C for 30 minutes and then spun at 13000rpm in a Heraeus A13 benchtop centrifuge for 15 minutes to precipitate the DNA. The supernatant was then poured off and the pellet allowed to air dry. The dry pellet was then resuspended in 150ul of water. Following this, 2 Chromospin-100 columns (Clonetech) were prepared for each sample by spinning the columns in a Hereaus 17RS centrifuge for 3 minutes at 3500rpm according to the manufacture's instructions. Following centrifugation 75ul of the DNA solution was added to each prepared column and spun as before collecting the purified DNA into a 1.5ml eppendorf tube. The 2 samples for each exon were then pooled and the DNA concentration measured by reading the absorption at 260nm and 280nm in a Pharmacia Genequant spectrophotometer.

## Solutions and Buffers

*1xTE pH7.6*

10mM Tris HCl

1mM EDTA

*TES pH7.5*

10mM Tris-HCl

1mM EDTA

2M NaCl

*1xFok I buffer pH7.9*

50mM potassium acetate

20mM Tris Acetate

10mM magnesium acetate

1mM DTT

*1xBbv I buffer Ph7.9*

50mM NaCl

10mM Tris-HCl

10mM MgCl2

1mM DTT

*1xSau 3A buffer pH7.9*

33mM Tris acetate

66mM potassium acetate

10mM magnesium acetate

0.5mM DTT

*1xLigase buffer pH7.8*

50mM Tris-HCl

10mM MgCl2

10mM DTT

1mM ATP

50ug/ml BSA

**Experiment**

*Concentrations of Column Purified DNA*

exon 14 - 130ng/ul

exon 16 - 120ng/ul

exon 19 - 115ng/ul

1ug exon14 (255bp) = 5.9pmol, 1ug exon16 (272bp) = 5.58pmol, 1ug exon19 (252bp) = 6.03pmol.
1ug exon14 = 7.7ul, 1ug exon16 = 8.3ul, 1ug exon19 = 8.7ul
therefore
exon 14 = 0.76 pmol/ul, exon 16 = 0.67pmol/ul, exon 19 = 0.69pmol/ul

Sau 3A Digest
30, 15 and 6pmol of column purified exons 14, 16 and 19, respectively, were digested with 20 units of Sau 3A in 100ul of 1xSau 3A buffer at 37°C for 4 hours.

| | |
|---|---|
| exon14 | 39.5ul |
| exon16 | 22.4ul |
| exon19 | 8.7ul |
| sau 3A | 5ul |
| 10xsau 3A buffer | 10ul |
| H2O | 14.4ul |

Following digestion the reaction mix was heated at 65oC in a Techne Dryblock for 20 minutes to inactivate the enzyme.

*Preparation of DynaBead M280*

According to the manufacture's instructions 3mg of DynaBeads M280 will bind 60-120 pmol of biotinylated double stranded DNA.

300ul of DynaBeads M280 at 1mg/ml were washed with 100ul TES by holding the beads to the side of an eppendorf tube with a Magnetic Particle Concentrator (Dynal UK) so that the supernatant could be removed. This was repeated three times (All subsequent bead manipulation were carried out in this manner according to manufacture's instructions). The beads were resuspended in 100ul of TES and the Sau 3A digested DNA added and incubated at room temperature for 1 hour to allow the biotinylated DNA to bind to the beads.

The Beads/DNA were then washed three times with 1xligase buffer using the Magnetic Particle Concentrator (Dynal UK) as before.

*Ligation of SauFAM Adaptor (Containing Fok I site)*

The supernatant was removed and the beads/DNA were resuspended in 75ul of 1xligase buffer containing 300pmol of SauFAM adaptor and 4000 units of ligase (New England Biolabs).
Beads/DNA ,7.5ul 10 ligase buffer, 15ul SauFAM (at 20pmol/ul), 10ul ligase (at400 units/ul), 42.5ul H20

The reaction was then incubated at 16°C for 2 hours

*Fok I Digestion*

Following ligation the beads/DNA were was 2 times with 75 ul of 1x Fok I buffer and the resuspended in 100ul of 1xFok I buffer and heated at 65oC in a Techne Dryblock for 20 minutes to inactivate any remaining ligase.

The buffer was was removed and the beads/DNA resuspended in 95ul of 1x Fok I buffer containing 20 units of Fok I (New England Biolabs)

Beads/DNA, 9.5ul 10x Fok I buffer, 5ul Fok I (at 4 units/ul)

The beads/DNA were then incubated at 37oC for 2 hours.

Following incubation the supernatant, containing the fragments cleaved by Fok I, was then transferred to a fresh eppendorf tube and heated at 65oC for 20 minutes in a Techne Dryblock in inactivate the Fok I.

*Ligation of Fok I Cleaved Fragments to Bbv Adaptors on Microtiter Plate*

The Fok I fragments were then divided into three tubes each containing 30ul of Fok I cleaved fragments, 5ul of 10x Ligase buffer, 3ul ligase (at 400uints/ul -New England Biolabs) and 12ul of H20.

The ligase buffer on a plate containing adaptors Bbv14, 16, 19 in separate wells (prepared as previously described) was removed and the above reaction mixtures, containing the Fok I cleaved fragments and ligase, added to each.

The wells were then incubated at 16°C for one hour and then washed three times with 50ul of TES. The TES was removed from the wells, another 50ul of TES added and the fluorescence measured in Biolumin Microplate reader (Molecular Dynamics). A well to which no fragments were added and just contained Bbv adaptors was used as a blank.

*Data expressed as RFUs*

| | |
|---|---|
| Bbv14 well | 1774 RFU |
| Bbv16 well | 1441 RFU |
| Bbv19 well | 1192 RFU |
| Blank | 1010 RFU |

The reading from the blank well, which is a background reading, was subtracted from the reading of the other wells and gave the following.

Bbv14 well 764 RFU
Bbv16 well 431 RFU
Bbv19 well 182 RFU

As half as much of exon 16 compared to exon 14 (15pmol exon 16, 30 pmol exon 14) was included into the procedure the reading obtained from the Bbv16 well should be half (i.e. 50%) of that obtained from the Bbv14 well and as one fifth the amount of exon 19 compared to exon 14 (6pmol exon 19, 30 pmol exon 14) the reading obtained from the Bbv19 well should be one fifth (i.e. 20%) that obtained from the Bbv14 well.

*Ideal Reading Expressed As Percentages*
Bbv14 well 100
Bbv16 well 50
Bbv19 well 20

*Actual Readings Expressed As Percentages (using Bbv14 well as 100%)*
Bbv14 well 100
Bbv16 well 56.4
Bbv19 well 23.8

Bbv16 well 6.4% error
Bbv19 well 3.8% error

Therefore, this process is capable of separating a mixed population of DNA ,and identifying 4bp, while at the same time maintaining the relative proportions of the original mixture with minimal errors. Which in turn can then be reprobed to obtain another 4bp and the associated quantitative data.

**Example 2**
**Ligation Optimisation**

The ligation reaction is a critical step in this sequencing technology. Therefore, full optimisation of this reaction is

required to ensure success with these techniques. The conditions for the ligation reaction have been investigated by ligating fluorescently (FAM) labelled adaptors to biotinylated adaptors captured to a streptavidin coated microtitre plate.

The biotinylated adaptors consist of a GC rich and AT rich type having the 4 base pair overhang sequence CGGC and TAAT respectively. These represent the extremes of GC and AT hybridisation and are therefore used to determine the conditions required to equalise their differing hybridisation kinetics.

Data (presented in RFUs - Relative Fluorescent Unit ) provided below are means of duplicate reactions. Reactions performed in 25ul.

*FAM labelled Adaptor Titration*

Used to evaluate the optimum amount of FAM labelled adaptor for use in the ligation reaction. To 1pmol of captured adaptor 0, 2.5, 5 and 10 pmol of FAM labelled adaptor was ligated to it for 5 minutes at 16°C.

| pmol GCCG | RFU | RFU | pmol ATTA |
|---|---|---|---|
| 0 | 56 | 71 | 0 |
| 2.5 | 175 | 614 | 2.5 |
| 5 | 303 | 756 | 5 |
| 10 | 293 | 783 | 10 |

The above data demonstrates that 5pmol of FAM labelled adaptor is optimum for this reaction. Increasing the amount further does little to promote further ligation. This experiment also demonstrates the ligation differences between GC and AT rich adaptors with GC rich ones ligating 2.5 times more that AT rich under identical conditions.

*Ligase Titration*

Used to evaluate optimum ligase concentration for the reaction. 5pmol of GC and AT rich adaptor was ligated to their appropriate captured adaptor for 5 minutes at 16°C with 0, 0.5, 1 and 2ul of ligase (at 4 units per ul - units as defined by New England Biolabs, not Weiss units).

| ul ligase GCCG | RFU | ul ligase ATTA | RFU |
|---|---|---|---|
| 0 | 95 | 133 | 0 |
| 0.5 | 186 | 355 | 0.5 |
| 1 | 205 | 420 | 1 |
| 2 | 224 | 383 | 2 |

These data suggest the 1ul of ligase is optimum for this reaction. Increasing the amount of enzyme in the reaction appears to have an inhibitory effect with the GC rich adaptor which maybe a result of the increased amount of glycerol in the reaction. The ligase is stored in a 50% glycerol solution.

Investigating the effect of a higher concentration ligase (also available from New England Biolabs) is proposed for future work. As this maybe a way equalising differences of GC and AT adaptors

by driving each reaction to completion.

*Reaction Time Course*

Investigation the effect of time on the reaction. 2.5 and 5pmol of GC and AT rich adaptor ligated to their appropriate captured adaptor for 5, 10, 30 and 60 minutes at 16°C.

| GCCG | 2.5 pmol | 5 pmol | ATTA | 2.5 pmol | 5 pmol |
|---|---|---|---|---|---|
| 5 min | 178 | 216 | 5 min | 97 | 123 |
| 10 min | 198 | 255 | 10 min | 107 | 148 |
| 30 min | 312 | 377 | 30 min | 229 | 216 |
| 60 min | 474 | 486 | 60 min | 231 | 326 |

As can be seen from these data increasing the reaction time increases the amount of FAM labelled adaptor ligated, as expected. A reaction time of 60 minutes will be impractical for the proposed techniques. However, these reactions do not contain any agents which promote ligation through intra molecular crowding such as polyethylene glycol (PEG) or ficol. By including such additives the reaction time can be reduced to an acceptable duration by increasing ligation speed.

*Titration of intra molecular crowders*

The intra molecular crowders PEG, ficol and hexamine chloride were titrated to investigate their effects on ligation. Tetremethly ammonium chloride, which modifies Watson and Crick base pairing, was also titrated to investigate its effect on the

differing efficiency of ligation of AT and GC rich adaptors. 5pmol of adaptor was ligated for 10 minutes at 16°C.

PEG Titration

| PEG % | GCGG | ATTA |
|---|---|---|
| 0 | 545 | 273 |
| 2.5 | 510 | 388 |
| 7.5 | 534 | 384 |
| 15 | 326 | 422 |

The addition of PEG to reaction appears to have little effect on the ligation of the GCCG adaptor up to 7.5% and at 15% it is inhibitory to the reaction. However, it increases the amount of the ATTA adaptor ligated. Further titration is required to determine at which concentration the efficiency of the two reactions is equal.

Ficol Titration

| ficol% | GCGG | ATTA |
|---|---|---|
| 0 | 545 | 273 |
| 2.5 | 550 | 341 |
| 7.5 | 570 | 398 |
| 15 | 274 | 152 |

As with PEG, the addition of ficol has little effect on the efficiency of the GCCG reaction up to 7.5% and is inhibitory at 15%. However, increasing the concentration of ficol increases the efficiency of the ATTA reaction up to 7.5% and at 15% it is inhibitory. Again, further titration is required to evaluate at which concentration the reactions are equalised.

Hexamine Chloride Titration

| mM hexamine chloride | GCGG | ATTA |
|---|---|---|
| 0 | 545 | 273 |
| 1 | 449 | 295 |
| 5 | 439 | 262 |
| 25 | 300 | 116 |

This data suggests that the addition of hexamine chloride in this system is inhibitory and therefore has little use in promoting ligation.

TMAC Titration

| mM TMAC | GCGG | ATTA |
|---|---|---|
| 0 | 545 | 273 |
| 1 | 681 | 453 |
| 5 | 647 | 215 |
| 25 | 686 | 97 |

The addition of TMAC to the reaction appears to increase the efficiency of the GCCG adaptor ligation at all concentrations tested, while decreasing the ATTA reaction at concentrations above 1mM. The inclusion of TMAC with PEG or ficol should be investigated as a means of equalising and promoting ligation of GCCG and ATTA adaptors.

It is clear that the inclusion of such intra molecular crowders, such as PEG or ficol, will allow the equalisation of the differences in reaction efficiency observed between AT and GC rich adaptors.

**Example 3**

**COMPETITIVE HYBRIDISATION ASSAY**

Additive Titration

It is important that differing efficiencies of ligation between AT and GC rich adaptors are reduced to a minimum if the ligation of adaptors is to be used to obtain quantitative data on a mixed population of nucleic acid.

The effects of polyethylene glycol 8000 and ficol on equalising the differing efficiencies of ligation between AT rich (ATTA) and GC rich (GCCG) adaptors has been investigated (see previous example). Results of further similar experiments are shown in Figure 9 (all results given as Relative Fluorescence Units (RFU)).

*Ficol Titration*

Figure 9 shows a graph representing the effect that increasing Ficol concentration has on the efficiency of ligating FAM

labelled GCCG adaptor (series 1) to captured CGGC target adaptor and FAM labelled ATTA adaptor (series 2) to captured TAAT target adaptor.

*RFU - Relative Fluorescent Unit*

As can be seen increasing the amount of ficol in the reaction mix increases the efficiency of reactions for the GC rich adaptor substantially and to a limited degree for the AT rich adaptor. At concentration above 10% the ficol is inhibitory.

Ficol has much less of an effect on the efficiency of these reactions as compared to PEG (see below) and therefore will be of less use in helping to equalise the efficiency of ligation between AT and GC rich adaptors.

*PEG Titration*

AT Rich reactions

Figure 9 also shows a graph representing the effect that increasing PEG concentration has on the efficiency of ligating FAM labelled ATTA adaptor to captured TAAT target adaptor.

Clearly increasing the concentration of PEG increases the amount of adaptor ligated to the target ,in a 5 minute reaction, up to around 10% when at concentration higher than this it begins to have an inhibitory effect. At 10% PEG concentration approximately 3 times more adaptor is successful ligated to the target than with no PEG in the reaction mix (1481 RFUs at 0%, 4369 RFUs at 10%).

This increasing in efficiency is probably due to the PEG increasing the time that the adaptor can hybridise to the target

therefore increasing the chance that the enzyme can ligate it to the target. At concentrations above 10% the reaction solution is rather viscous and this will decrease the mobility of the reaction components and hence reduce reaction efficiency i.e. the benefits of having the adaptor hybridised to the target for long is lost if reaction components cannot get to the adaptor quick enough to complete the reaction. This would explain the inhibitory effect at the higher PEG concentrations.

*GC Rich Reactions*

Figure 9 also shows a graph representing the effect that increasing PEG concentration has on the efficiency of ligating FAM labelled GCCG adaptor to a captured CGGC target adaptor.

Interestingly, increasing the concentration of PEG in this reaction has a general inhibitory effect. This observation is probably due to the increased viscosity of the solution reducing the mobility of the reaction components and therefore the reaction efficiency. This appears to out way any effects that increased hybridisation times may have on the efficiency of the reaction. This is probably due to the fact that GC rich adaptors hybridise more strongly (as compared to AT rich ones) due to the extra hydrogen bond that GC base pairs have and increasing the concentration of PEG does little to increase the time that the adaptors remain hybridised.

Therefore, from this data, it is proposed that a PEG concentration of 10% should be used in a reaction where the ligation of adaptors is used to obtain quantitative data.

*Competitive assays*

If one has 256 uniquely mass labelled adaptors one reduce the time it take to obtain quantitative data from a mixed population

of nucleic acid by ligating 128 adaptors simultaneously followed by 128 of the corresponding complimentary adaptors. In this system adaptors would compete with each other for their complimentary sites. In order to investigate the specificity of this system various FAM labelled adaptors (specific e.g. GCCG and to CGGC mismatched e.g. GCGG to CGGC) have been tested against increasing concentrations of unlabelled specific adaptors under different conditions e.g. differing enzyme concentrations. The following are the data from conditions which produced the greatest specificity and reproducibility. All reactions were performed in duplicate.

*AT Rich Adaptors*

Firstly, 2.5 pmol FAM labelled specific (ATTA) and mismatched (ATCA and ATAA) were ligated in the presence of 0,1.25,2.5 and 5 pmol of unlabelled ATTA with 10% PEG for 5 minutes at 16oC. Following incubation unligated adaptors were removed by 3 washes of 1xTE and the fluorescence measured. Data are expressed as relative fluorescent units.

| | 2.5 pmol FAM-ATTA | 2.5 pmol FAM-ATCA | 2.5 pmol FAM-ATAA |
|---|---|---|---|
| 0 pmol ATTA | 3285 | 0 | 223 |
| 1.25 pmol ATTA | 2994 | 0 | 0 |
| 2.5 pmol ATTA | 2744 | 0 | 0 |
| 5 pmol ATTA | 1605 | 0 | 0 |

The most important observation from the above data is that the ATCA mismatched adaptor does not ligate to any measurable degree. The presence of the C in the ATCA adaptor must therefore disrupt the base pairing completely thereby preventing any ligation. One would predict that this would also be the case if the mismatch contained a G instead of a C. The ATAA adaptor only ligates at 6.7% of the amount as the ATTA adaptor. The replacement of the T with an A in this mismatch therefore

disrupts base pairing to a lesser degree than a C and therefore allows some ligation. However, the ligation of this mismatched adaptor is completely displaced by the presence of any unlabelled specific ATTA adaptor.

From these data one can therefore conclude that the ligation of AT rich adaptor will be highly specific in a competitive system and will deliver highly representative quantitative data.

*GC Rich Adaptors*

2.5 pmol FAM labelled specific (GCCG) and mismatched (GCAG and GCGG) were ligated in the presence of 0,1.25,2.5 and 5 pmol of unlabelled GCCG with 10% PEG for 5 minutes at 16oC. Following incubation unligated adaptors were removed by 3 washes of 1xTE and the fluorescence measured.
Data are expressed as relative fluorescent units.

| | 2.5 pmol FAM-GCCG | 2.5 pmol FAM-GCAG | 2.5 pmol FAM-GCGG |
|---|---|---|---|
| 0 pmol GCCG | 8615 | 2992 | 3311 |
| 1.25 pmol GCCG | 5091 | 2442 | 1472 |
| 2.5 pmol GCCG | 3660 | 1991 | 841 |
| 5 pmol GCCG | 2430 | 501 | 267 |

With GC rich sequences the mismatched adaptors do ligate as compared to the AT rich one which do not. The GCAG mismatch ligates at 35% and the GCGG at 38% of the amount of the specific GCCG. This is to be expected as GC base pairing is stronger than AT base pairing and can thus accommodate a degree of base pairing disruption. However, when in competition with equal amounts of unlabelled specific GCCG adaptor the amount of ligation achieved is reduced to 23% for the GCAG and 10% for the GCGG adaptors.

Therefore, the above data suggest that the ATTA rich adaptors will be highly specific but 10 to 23% of the GC rich adaptors

could ligate to an incorrect sequence. However, this should not be a problem as these errors can be compensated for in the software that would be required to analyse the data .Also, if the same experiment is repeated with a different reference enzyme one could use each set of data to cross reference the quantification and sequence data in order to resolve discrepancies produced from non-specific ligations.

**Adaptor Sequences and Preparation (Examples 2 and 3):**

ATTA Adaptor:

5'-FAM-GCATCAGGATGTACAG-3'

3'-CGTAGTCCTACATGTCATTA-PO4-5'

ATAA Adaptor:

5'-FAM-GCATCAGGATGTACAG-3'

3'-CGTAGTCCTACATGTCATAA-PO4-5'

ATGA Adaptor:

5'-FAM-GCATCAGGATGTACAG

3'-CGTAGTCCTACATGTCATGA-PO4-5'

GCCG Adaptor:

5'-FAM-GCATCAGGATGTACAG

3'-CGTAGTCCTACATGTCGCCG-PO4-5'

GGCG Adaptor:

5'-FAM-GCATCAGGATGTACAG

3'-CGTAGTCCTACATGTCGGCG-PO4-5'

GACG Adaptor:

5'-FAM'GCATCAGGATGTACAG-3'

3'-CGTAGTCCTACATGTCGACG-PO4-5'

TAAT Adaptor:

5'-Biotin-GCATCAGGATGTACAG-3'

3'-CGTAGTCCTACATGTCTAAT-PO4-5'

CGGC Adaptor:

5'-Biotin-GCATCAGGATGTACAG-3'

3'-CGTAGTCCTACATGTCCGGC-PO4-5'

Adaptors prepared as described in Example 1.

Abbreviations:

FAM - fluorescein

PO4 - phosphate

All oligonucleotides purchased from Oswel DNA services.

**Reconstructing sequences from matrices of n-mers:**

A sequencing reaction by this method involves repeated cycles of cleaving a template with a type IIs restriction endonuclease whose recognition sequence is provided by an adaptor. If the reaction is peformed with multiple templates then each cycle of the sequencing reaction will generate a signal for a series of n-mers. Many cycles will of the reaction will generate a matrix of n-mers which must be analysed to reconstruct the sequences of the source templates.

An algorithm has been implemented in the C programming language that can import and interpret such a data matrix. The entire program is not listed but the critical data structure to store the n-mer matrix is shown with the critical sections of code. A complete program that can form part of a data capture and processing system should be trivial to develop from this starting point.

The program operates by first analysing the data matrix to identify in each column of the matrix, corresponding to one cycle of the sequencing reaction, n-mer frequencies or quantities which are equivalent in other columns of the matrix given a predefined margin of error in the measurement of n-mer quantities within which to operate. The raw n-mer frequencies in the data matrix are then replaced with their probable group frequencies in each

column.

This new data matrix is then analysed by a second algorithm which assumes that there should be the same number of n-mers in each column of the matrix and attempts to resolve any 'sums' of frequencies where the same n-mer has occurred in more than one template in a given cycle of the sequencing reaction. This algorithm takes the group frequencies in the data matrix and generates a sorted 'frequency list' that lists the number of occurrences of each group frequency in order of increasing number of occurrences.

The algorithm then takes group frequencies with the lowest number of occurrences first on the assumption that these are likely to be sums, since sums of groups should occur with a relatively low frequency. An alternative would be to generate a sorted list of group frequencies, in order of decreasing quantity, and start with the largest quantities, again on the assumption that these are likely to be sums.

The algorithm then tests each frequency in the list against each column of the original data matrix. If the group frequency occurs in the column it is tested against all combinations of pairs of group frequencies that are missing from the column to see if any of these missing frequencies can add up to give the current frequency being tested. If any of these missing frequencies do add up and there is only one pair that can add up within the predetermined margin of error then it is assumed that the larger frequency is the sum of the two missing frequencies and the larger frequency is replaced in the current column of the data matrix by occurrences of the two missing frequencies. Any frequencies are the sum of two pairs of missing frequencies are marked as such and in the final sequence reconstruction the bases are marked as unknown.

At the end of this analysis one should be left, in most cases, with a data matrix that has the same group frequencies in each

column of the matrix. Each group frequency is then assumed to correspond to a single template and the sequence of each template is then the series of n-mers from all the columns in the matrix identified by its group frequency.

*C Header File of N-mer Matrix Data Structure:*

```
#include <stdio.h>
#include <stdlib.h>

#include <string.h>

#include <math.h>

#include "List.h"


#define CYCLES 10

#define NMERS 256

#define CUTLENGTH 4



typedef struct

{

      double element[4];


} Element;


typedef struct

{

      Element matrix[NMERS][CYCLES];


      List *frequencyList;


} SeqMatrix;
```

```
void InitSeqMatrix(SeqMatrix *myMatrix);

void AddtoMatrix(SeqMatrix *myMatrix, int columnNo, int fourMer, double frequency);

int ReplaceElement(SeqMatrix *myMatrix, double oldFrequency, double newFrequency);

int FindColumn(SeqMatrix *myMatrix, double frequency);

int InColumn(SeqMatrix *myMatrix, double frequency, int columnNo);

int FindRow(SeqMatrix *myMatrix, double frequency, int columnNo);

void PrintMatrix(SeqMatrix *myMatrix);

void FPrintMatrix(SeqMatrix *myMatrix);

void BuildList(SeqMatrix *myMatrix);

void ResolveGroups(SeqMatrix *myMatrix, int error);

void ResolveSums(SeqMatrix *myMatrix, int errorSize);

void Reconstruct(SeqMatrix *myMatrix);
```

The SeqMatrix data structure stores the matrix of n-mers generated by a sequencing reaction.

The critical algorithms are:

° `ResolveGroups(SeqMatrix *myMatrix, int error)` which analyses the SeqMatrix data structure to identify frequencies in each column of the

matrix that are equivalent.

- The `ResolveSums(SeqMatrix *myMatrix, int errorSize)` algorithm analyses the matrix generated by the ResolveGroups algorithm and attempts to determine which frequencies are sums of other frequencies.

- The `Reconstruct(SeqMatrix *myMatrix)` algorithm analyses the data matrix produced by the ResolveSums algorithm to reconstruct the sequences encoded by the group frequencies (not listed).

The program as it stands is far from optimal but will reliably reconstruct model matrices generated from fifteen template simultaneously with an error in the measurement of frequencies of about 2%. Many improvements to the basic algorithms can still be made.

To be useful in analysing real data, noise subtraction algorithms would be needed and an algorithm to normalise frequencies in each column to account for progressive decrease in signal with each cycle of the sequencing reaction that will result from the fact that no enzymatic step will be 100% efficient.

*ResolveGroups code:*

```
// attempts to group error laden frequencies into groups corresponding to
// original source frequencies

void ResolveGroups(SeqMatrix *myMatrix, int error)
{
    int i, j, k, l, n, groupCount, ambiguity;
    double maxDistance, minDistance, localError, fError, meanFreq, groups[400][CYCLES];
    List *tempList = NULL, *newList = NULL;


    printf("\nResolving groups of frequencies...\n");

    for(i = 0; i < 100; i++)
    {
        for(j = 0; j < CYCLES; j++)
        {
            groups[i][j] = 0.0;
        }
    }

    myMatrix->frequencyList = FreqSortList(myMatrix->frequencyList);

    // twice percent error as first grouping attempt
    fError = (double)(2 * error);

    // test frequencies to see if they belong to a group
    tempList = myMatrix->frequencyList;
    i=0;
    while(tempList != NULL)
    {
        localError = (fError/100 * tempList->frequency);

        // value that is 2% greater than first frequency -
        maxDistance = tempList->frequency + localError;

        // value that is 2% less than first frequency -
        minDistance = tempList->frequency - localError;

        for(j=0;j<CYCLES;j++)
            {
                ambiguity = 0;
                for(k=0;k<NMERS;k++)
                {
                    // is 2nd frequency within twice percent error of 1st
                    // frequency?
                    if(myMatrix->matrix[k][j].element[0] > minDistance)
                    {
                        if(myMatrix->matrix[k][j].element[0] < maxDistance)
                        {
                            groups[i][j] =
                                myMatrix->matrix[k][j].element[0];
                            ambiguity += 1;
                        }
```

```
                        }
                    }
                    if(ambiguity >1)
                    {
                        printf("ambiguous grouping - first grouping\n");
                        groups[i][j] = -1.0;
                    }
                }
            tempList = tempList->next;
            i++;
}

/ *
// print out groups matrix to screen
for(k=0;k<i;k++)
{
        for(j=0;j<CYCLES;j++)
        {
            if(groups[k][j] != 0)
            {
                printf("%5.4d ",(int)groups[k][j]);
            }
        }
        printf("\n");
}
* /

// To hold new list of grouped frequencies
newList = NULL;

meanFreq = 0;
groupCount = 0;

for(j=0;j<CYCLES;j++)
{
        if(groups[0][j] != 0 && groups[0][j] != -1.0)
        {
            meanFreq += groups[0][j];
            groupCount += 1;
        }
}
meanFreq = meanFreq/groupCount;
//printf("%f, %d\n", meanFreq, groupCount);

k = 0;
l = 0;
while(k<i && l<i)
{
        for(l=k+1;l<i;l++)
        {
            for(j=0;j<CYCLES;j++)
            {
                if(groups[k][j] != groups[l][j])
                {
                    newList = AppendElement(newList, NewElement(meanFreq,
                groupCount));
```

```
                        k = l;
                        groupCount = 0;
                        meanFreq = 0;
                        for(n=0;n<CYCLES;n++)
                        {
                                if(groups[k][n] != 0 && groups[k][n] != -1.0)
                                {
                                        meanFreq += groups[k][n];
                                        groupCount += 1;
                                }
                        }
                        meanFreq = meanFreq/groupCount;
                        //printf("%f, %d\n", meanFreq, groupCount);
                }
        }
    }
}

newList = AppendElement(newList, NewElement(meanFreq, groupCount));

// narrow down allowed error range as second grouping attempt:
fError = (double)(1.5 * error);

// Replace error ridden frequencies with corresponding group mean values
// retesting each value against group means

tempList = newList;
while (tempList != NULL && tempList->frequency != 0)
{
        localError = (fError/100 * tempList->frequency);

        // value that is 2% greater than first frequency -
        maxDistance = tempList->frequency + localError;

        // value that is 2% less than first frequency -
        minDistance = tempList->frequency - localError;
        for(l=0;l<CYCLES;l++)
        {
            ambiguity = 0;
            for(n=0;n<NMERS;n++)
            {
                    // is 2nd frequency within twice percent error of 1st frequency?
                    if(myMatrix->matrix[n][l].element[0] > minDistance)
                    {
                            if(myMatrix->matrix[n][l].element[0] < maxDistance)
                            {
                                    myMatrix->matrix[n][l].element[0] =
                                            tempList->frequency;
                                    ambiguity += 1;
                            }
                    }
            }
            if(ambiguity >1)
            {
                    printf("ambiguous grouping - second grouping\n");
                    groups[i][j] = -1;
```

```
            }
        }
        tempList = tempList->next;
    }

    // Update frequencyList with values of group means
    myMatrix->frequencyList = InitList(myMatrix->frequencyList);
    myMatrix->frequencyList = CopyList(newList);
}
```

*ResolveSums code:*

```
// Resolve frequencies that are the sums of atomic quantities

void ResolveSums(SeqMatrix *myMatrix, int errorSize)
{
   int flag, colNo, rowNo, candidates, i, j, k;
   double minSum, maxSum, tempFreq, fError, localError;
   List *tempList1, *tempList2, *tempList3, *tempList4, *tempList5;

   printf("\nResolving sums of frequencies...\n");

    myMatrix->frequencyList = SortList(myMatrix->frequencyList);

   fError = ((double)errorSize);
   tempList1 = myMatrix->frequencyList;
  while(tempList1 != NULL)
  {
       flag =1;
       localError = (fError/100 * tempList1->frequency);

       // value that is within % error greater than first frequency -
       maxSum = tempList1->frequency + localError;

       // value that is within % error less than first frequency -
       minSum = tempList1->frequency - localError;

       while(tempList1->occurrences>0 && flag)
       {
         // FindColumn returns the column in which the frequency is found

          colNo = FindColumn(myMatrix, tempList1->frequency);
             tempList2 = tempList1->next;
             candidates = 0;
             while(tempList2 != NULL)
             {
                if(!(InColumn(myMatrix, tempList2->frequency, colNo)))
                {
                  tempList3 = tempList2->next;
                  while(tempList3 != NULL)
                  {
                        if(!(InColumn(myMatrix, tempList3->frequency, colNo)))
                        {
                          tempFreq = tempList2->frequency;
                          tempFreq += tempList3->frequency;

                          if(minSum <= tempFreq)
                          {
                                if(maxSum >= tempFreq)
                                {
                                   printf("Frequency %f ",tempList1->frequency);
                                  printf("could be composed of %f", tempList2->frequency);
                                   printf("and %f\n",tempList3->frequency);
                                   SubtractOccurrence(myMatrix->frequencyList,
                                        tempList1->frequency);
```

```
            rowNo = FindRow(myMatrix, tempList1->frequency,colNo);
            myMatrix->matrix[rowNo][colNo].element[0]
                = tempList2->frequency;
            myMatrix->matrix[rowNo][colNo].element[1]
                = tempList3->frequency;
            AddOccurrence(myMatrix->frequencyList,
                tempList2->frequency);
            AddOccurrence(myMatrix->frequencyList,
                tempList3->frequency);
          candidates++;
        }
    }
}
tempList4 = tempList3->next;
while(tempList4 != NULL)
{
  if(!(InColumn(myMatrix, tempList4->frequency, colNo)))
  {
      tempFreq = tempList2->frequency;
      tempFreq += tempList3->frequency;
      tempFreq += tempList4->frequency;

      if(minSum <= tempFreq)
      {
        if(maxSum >= tempFreq)
        {
            printf("Frequency %f ",tempList1->frequency);
            printf("could be composed of %f ",
                tempList2->frequency);
            printf("and %f\n", tempList3->frequency);
            printf("and %f\n", tempList4->frequency);
            SubtractOccurrence(myMatrix->frequencyList,
                tempList1->frequency);
            rowNo = FindRow(myMatrix, tempList1->frequency,colNo);
            myMatrix->matrix[rowNo][colNo].element[0] =
                tempList2->frequency;
            myMatrix->matrix[rowNo][colNo].element[1] =
                tempList3->frequency;
            myMatrix->matrix[rowNo][colNo].element[1] =
                tempList4->frequency;
            AddOccurrence(myMatrix->frequencyList,
                tempList2->frequency);
            AddOccurrence(myMatrix->frequencyList,
                tempList3->frequency);
            AddOccurrence(myMatrix->frequencyList,
                tempList4->frequency);
            candidates++;
        }
      }
}
tempList5 = tempList4->next;
while(tempList5 != NULL)
{
    if(!(InColumn(myMatrix, tempList5->frequency, colNo)))
    {
      tempFreq = tempList2->frequency;
```

```
                        tempFreq += tempList3->frequency;
                        tempFreq += tempList4->frequency;
                        tempFreq += tempList5->frequency;

                        if(minSum <= tempFreq)
                        {
                              if(maxSum >= tempFreq)
                              {
                                 printf("Frequency %f ", tempList1->frequency);
                                 printf("could be composed of %f ",
                                       tempList2->frequency);
                                 printf("and %f\n",tempList3->frequency);
                                 printf("and %f\n",tempList4->frequency);
                                 printf("and %f\n",tempList5->frequency);
                                 SubtractOccurrence(
                                       myMatrix->frequencyList,tempList1->frequency);
                                 rowNo = FindRow(myMatrix,
                                       tempList1->frequency,colNo);
                                 myMatrix->matrix[rowNo][colNo].element[0] =
                              tempList2->frequency;
                                 myMatrix->matrix[rowNo][colNo].element[1] =
                              tempList3->frequency;
                                 myMatrix->matrix[rowNo][colNo].element[1] =
                              tempList4->frequency;
                                 myMatrix->matrix[rowNo][colNo].element[1] =
                              tempList5->frequency;
                                 AddOccurrence(myMatrix->frequencyList
                                       tempList2->frequency);
                                 AddOccurrence(myMatrix->frequencyList,
                                       tempList3->frequency);
                                 AddOccurrence(myMatrix->frequencyList,
                                       tempList4->frequency);
                                 AddOccurrence(myMatrix->frequencyList,
                                       tempList5->frequency);
                                 candidates++;
                              }
                           }
                        }
                        tempList5 = tempList5->next;
                     }
                     tempList4 = tempList4->next;
                  }
                  tempList3 = tempList3->next;
               }
            }
            tempList2 = tempList2->next;
         }

         if(candidates > 1)
         {
            flag = 0;
              printf("Ambiguity - candidates = %d :",candidates);
              printf("Frequency = %f\n",tempList1->frequency);
         }

}
```

```
        tempList1 = tempList1->next;
    }
    myMatrix->frequencyList = RemoveNullOcc(myMatrix->frequencyList);
}
```

Key for Drawings

FIGURE 4

Step 1 Cleave genomic DNA with type IIs restriction endonuclease

Step 2 Add adaptors to fragments each bearing primer binding sites such that each sticky-end or subset thereof bears a unique primer site

Step 3 Differentially amplify by PCR by adding different amounts of primer for each adaptor

FIGURE 5

Step 1 Cleave genomic DNA with type IIs restriction endonuclease

Step 2 Ligate adaptor pair to fragments to tag termini

Step 3 Capture fragments to allow fragments with adaptor 2 at both termini to be washed away

Step 4 Cleave adaptor 2 with restriction endonuclease

Step 5 Release fragments from solid phase substrate

Step 6 Ligate capture adaptor to blunt end generated from fragments with adaptor 2 at one end

Step 7 Capture fragments or perform arbitrary further sorting

FIGURE 8

| | |
|---|---|
| Sorting step | Sort fragments onto array of oligonucleotides or into array of 256 wells |
| Cleavage step | Cleave immobilised fragments with type IIs restriction endonuclease corresponding to directionality adaptor 1 |
| Addition of Adaptor | Add adaptor with fluorescent label and with sticky-end complementary to one of the 256 possible 4 base overlaps that might be present on the immobilised nucleic acid fragments. The sequence of each adaptor's sticky-end must be known |

CLAIMS:

1. A method for sequencing nucleic acid, which comprises:

(a) obtaining a target nucleic acid population comprising nucleic acid fragments in which each fragment is present in a unique amount and bears at one end a sticky end sequence of predetermined length and unknown sequence,

(b) protecting the other end of each fragment, and

(c) sequencing each of the fragments by

(i) contacting the fragments with an array of adaptor oligonucleotides under hybridisation conditions, each adaptor oligonucleotide bearing a label, a sequencing enzyme recognition site, and a known unique base sequence of same predetermined length as the sticky end sequence, the array containing all possible base sequences of that predetermined length; removing any unhybridised adaptor oligonucleotide and recording the quantity of any hybridised adaptor oligonucleotide by detection of the label, then repeating the cycle, until all of the adaptors in the array have been tested;

(ii) contacting the hybridised adaptor oligonucleotides with a sequencing enzyme which binds to the recognition site and cuts the fragment to expose a new sticky end sequence which is contiguous with or overlaps the previous sticky end sequence;

(iii) repeating steps (i) and (ii) for a sufficient number of times and determining the sequence of the fragment by comparing the quantities recorded for each sticky end sequence.

2. A method according to claim 1, wherein each label comprises a mass label associated with a corresponding known base sequence for identifying the corresponding base sequence in mass spectrometry.

3. A method according to claim 2, wherein each adapter oligonucleotide labelled with an associated mass label is uniquely resolvable in mass spectrometry from the other labelled adapter oligonucleotides.

4. A method according to claim 3, wherein each adapter oligonucleotide is composed of nucleotide analogues which are resistant to fragmentation in the mass spectrometer.

5. A method according to claim 2, wherein each mass label is cleavably attached to its corresponding adaptor oligonucleotide and uniquely resolvable in mass spectrometry.

6. A method according to any one of claims 2 to 5, wherein the mass spectrometry is effected using a mass spectrometer with orthogonal time of flight or array detector geometry.

7. A method according to any one of the preceding claims, wherein the fragments are contacted in step (i) with the array of adaptor oligonucleotides in a cycle wherein the cycle comprises sequentially contacting each adaptor oligonucleotide of the array with the fragments.

8. A method according to any one of the preceding claims, wherein the target nucleic acid population is subjected to a step of sorting into sub-populations according to their sticky end

sequences and each of the sub-populations is subjected to steps (b) and (c).

9. A method according to any one of the preceding claims, wherein each fragment is produced by differential amplification.

10. A method according to any one of the preceding claims, wherein the predetermined length of the base sequence of the sticky ends is from 1 to 5.

11. A method according to any one of the preceding claims, wherein the sequencing enzyme comprises a type IIs restriction endonuclease.

12. A method according to any one of the preceding claims, wherein the target nucleic acid population comprises heterogenous nucleic acid fragments.

13. A method according to any one of the preceding claims, wherein the other end of each fragment is protected by ligation with an immobilisation adaptor oligonucleotide.

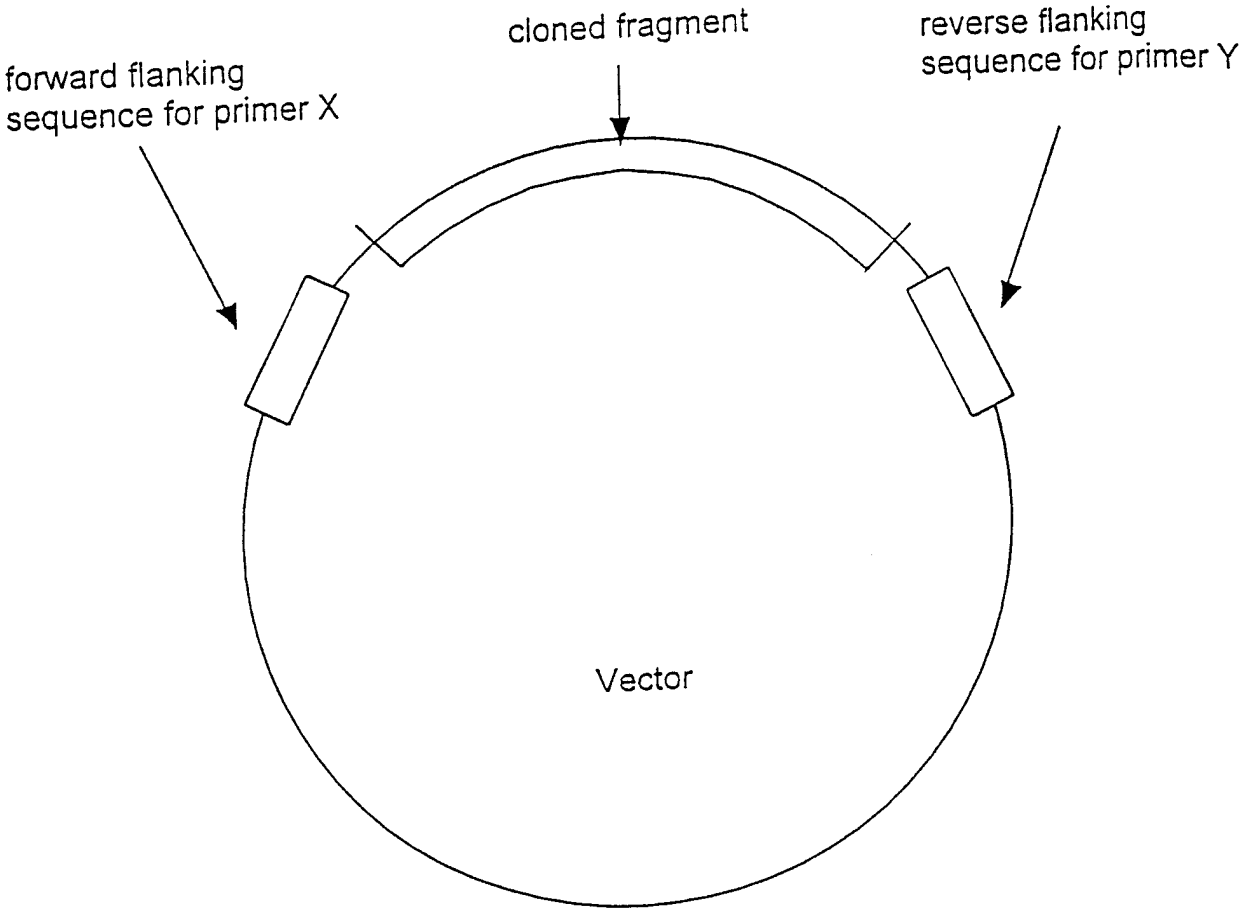

cloned fragment
reverse flanking
sequence for primer Y
forward flanking
sequence for primer X
Vector

FIGURE 1

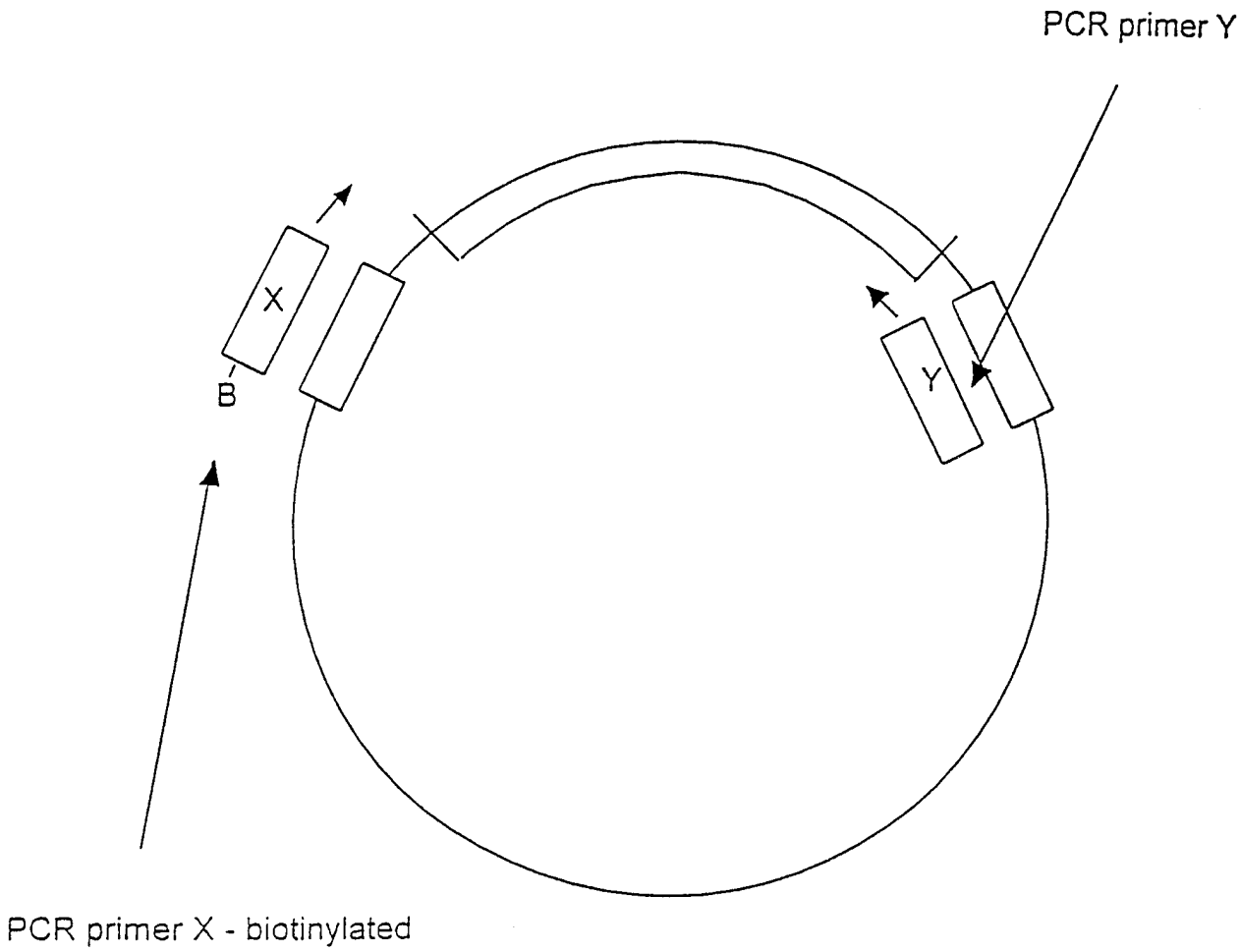
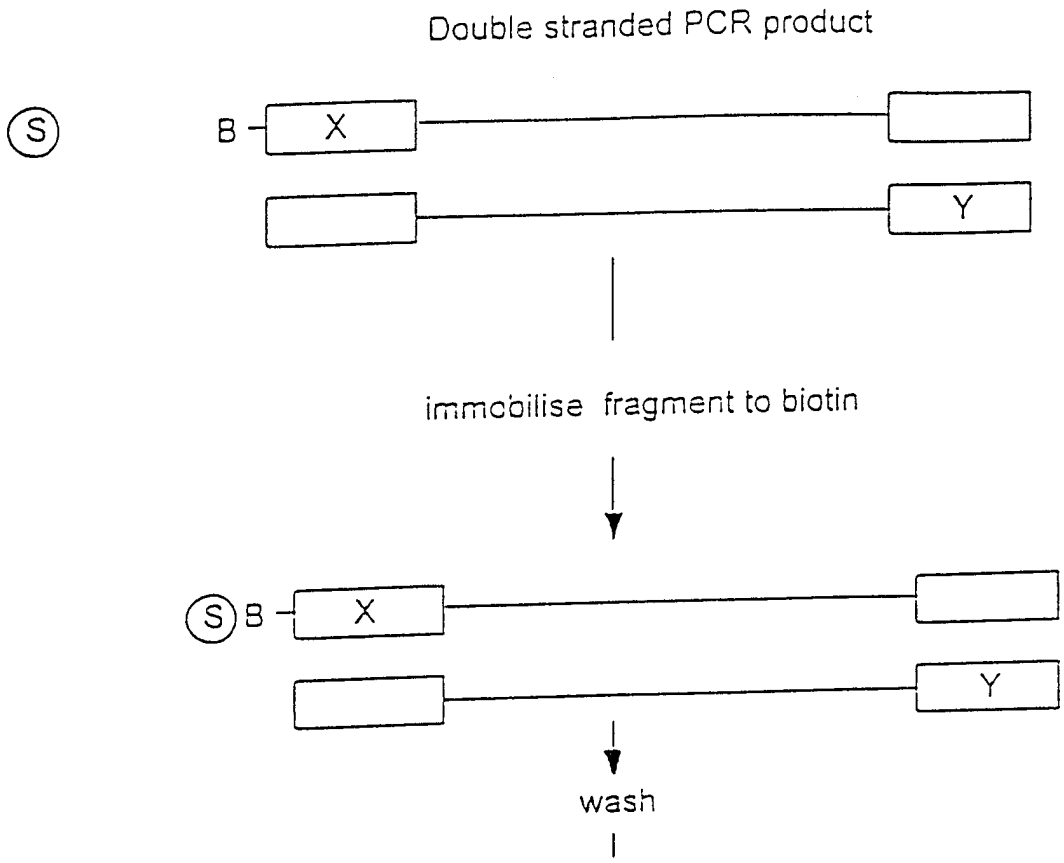

PCR primer Y
X
B
Y
PCR primer X - biotinylated
PCR amplify
Double stranded PCR product
B
X
Y
B - Biotin

FIGURE 2




S - Streptavidin coated bead or well
B - Biotin
x - primer x sequence
y - primer y sequence

FIGURE 3

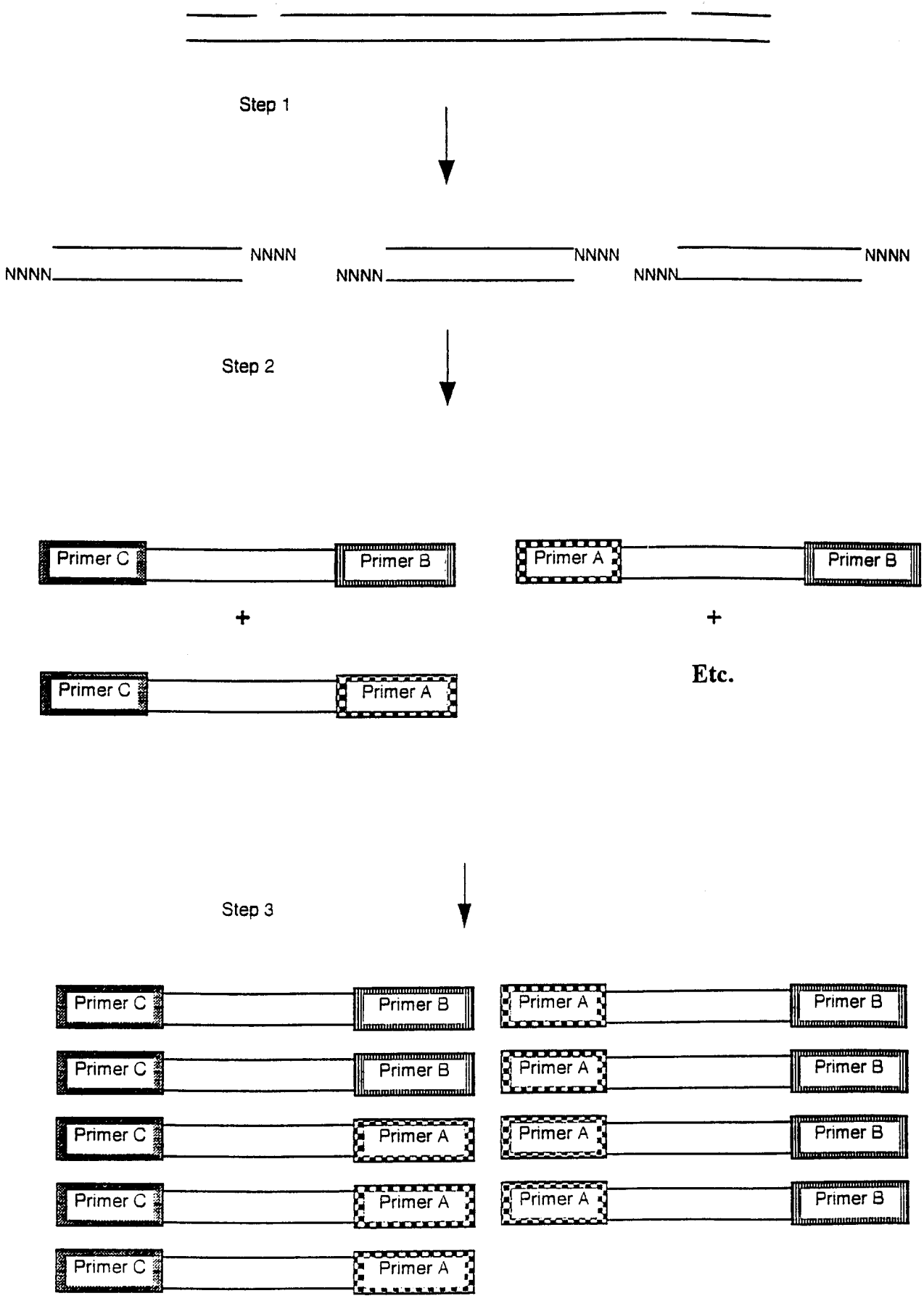

Step 1
NNNN
NNNN
NNNN
NNNN
NNNN
NNNN
Step 2
Primer C
Primer B
+
Primer C
Primer A
Primer A
Primer B
+
Etc.
Step 3
Primer C
Primer B
Primer C
Primer B
Primer C
Primer A
Primer C
Primer A
Primer C
Primer A
Primer A
Primer B
Primer A
Primer B
Primer A
Primer B
Primer A
Primer B

FIGURE 4

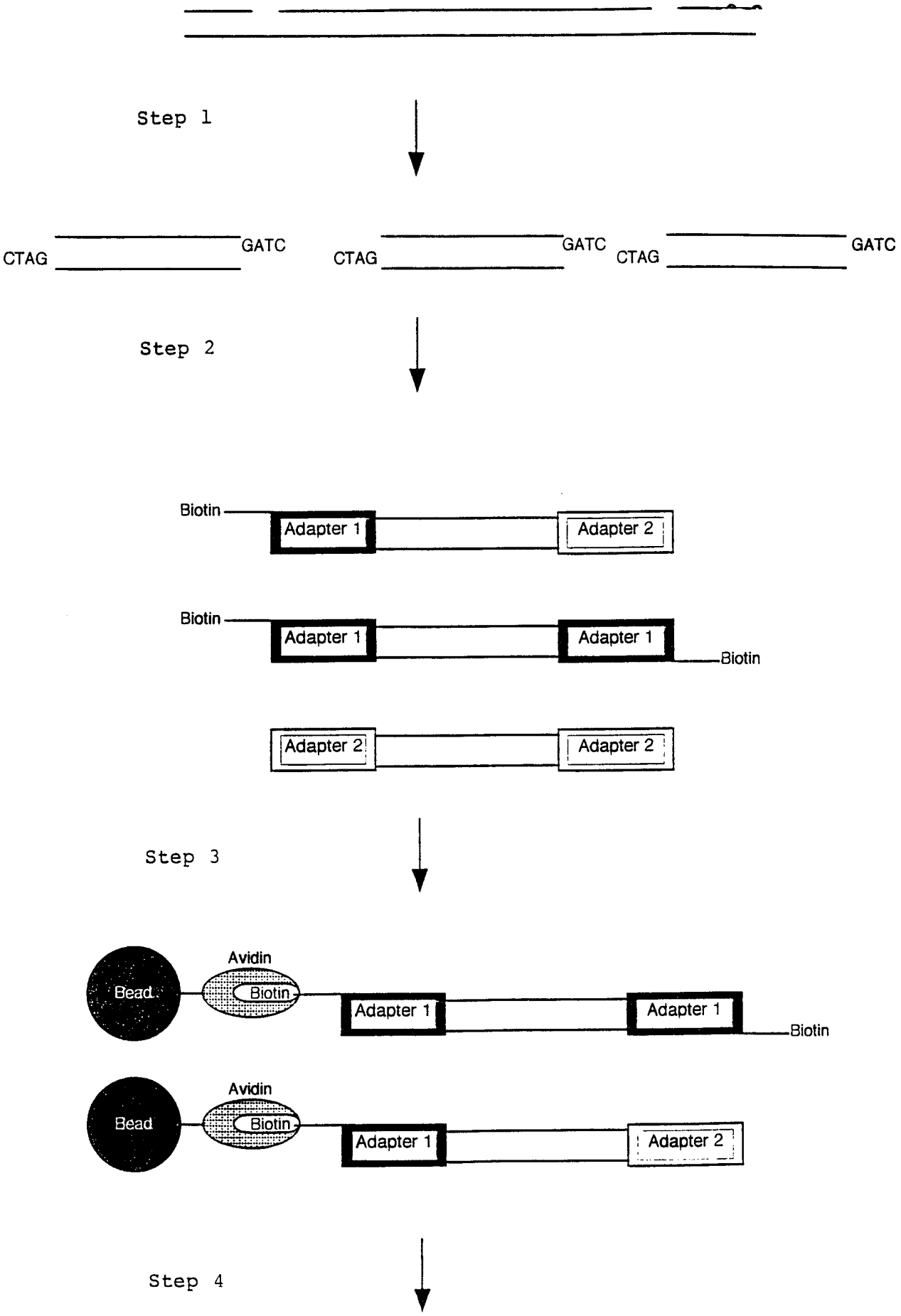

Step 1
CTAG
GATC
CTAG
GATC
CTAG
GATC
Step 2
Biotin
Adapter 1
Adapter 2
Biotin
Adapter 1
Adapter 1
Biotin
Adapter 2
Adapter 2
Step 3
Bead
Avidin
Biotin
Adapter 1
Adapter 1
Biotin
Bead
Avidin
Biotin
Adapter 1
Adapter 2
Step 4

FIGURE 5

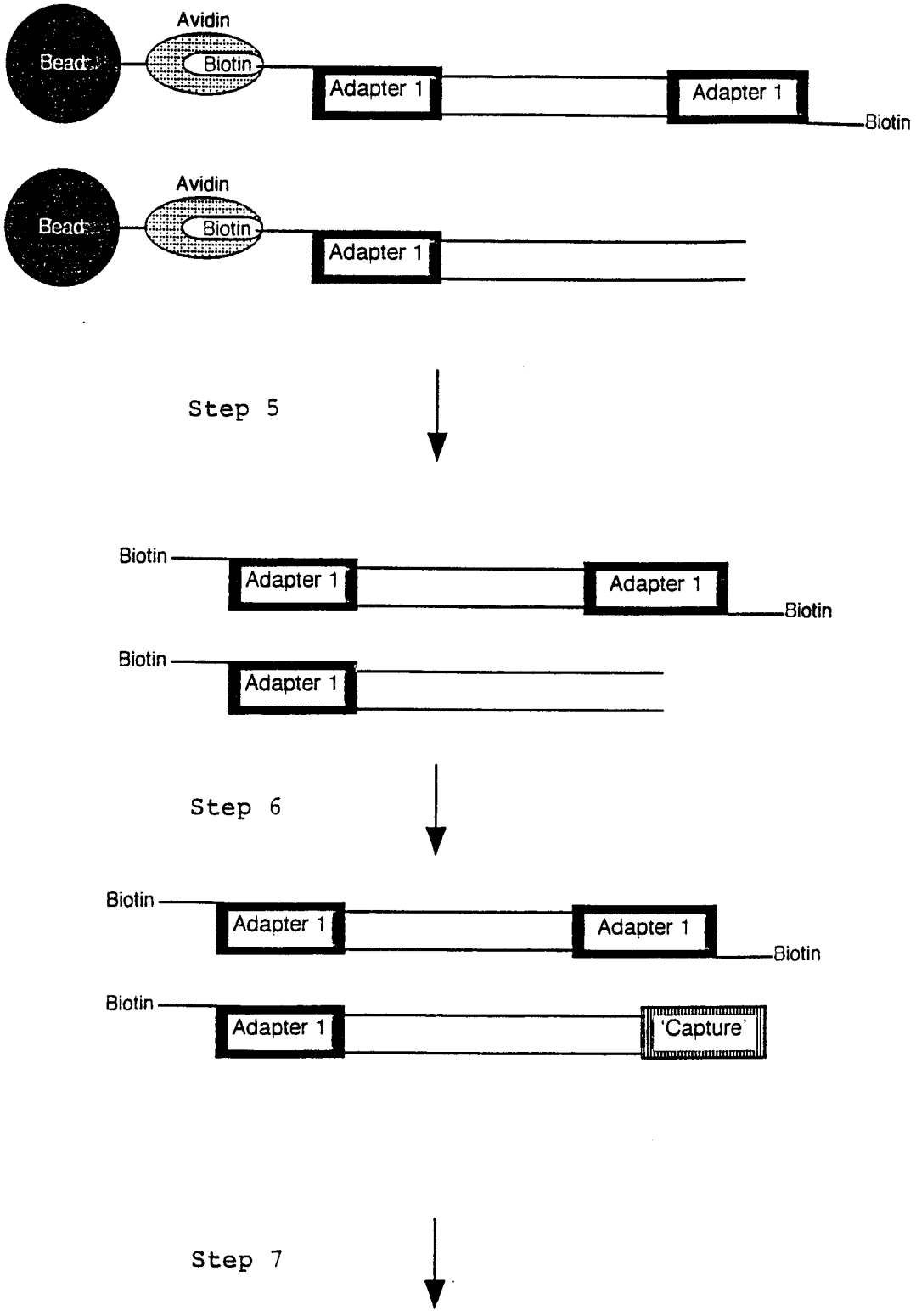



FIGURE 5 (Continued)

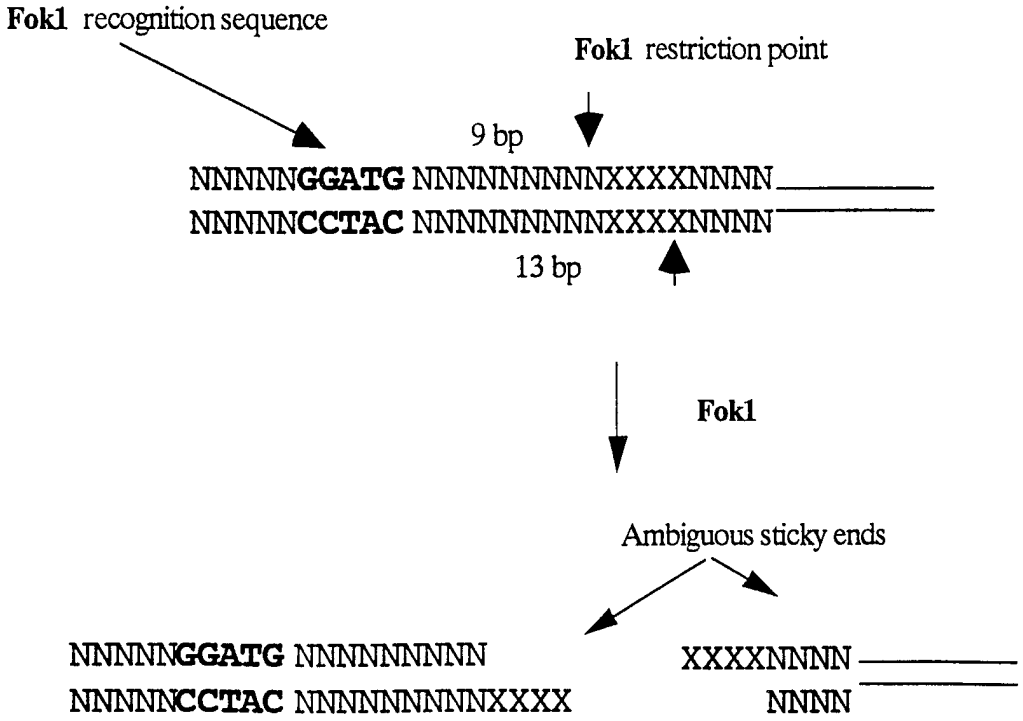

FokI recognition sequence
FokI restriction point
9 bp
NNNNNGGATG NNNNNNNNNXXXXNNNN
NNNNNCCTAC NNNNNNNNNXXXXNNNN
13 bp
FokI
Ambiguous sticky ends
NNNNNGGATG NNNNNNNNN
NNNNNCCTAC NNNNNNNNNXXXX
XXXXNNNN
NNNN

FIGURE 6

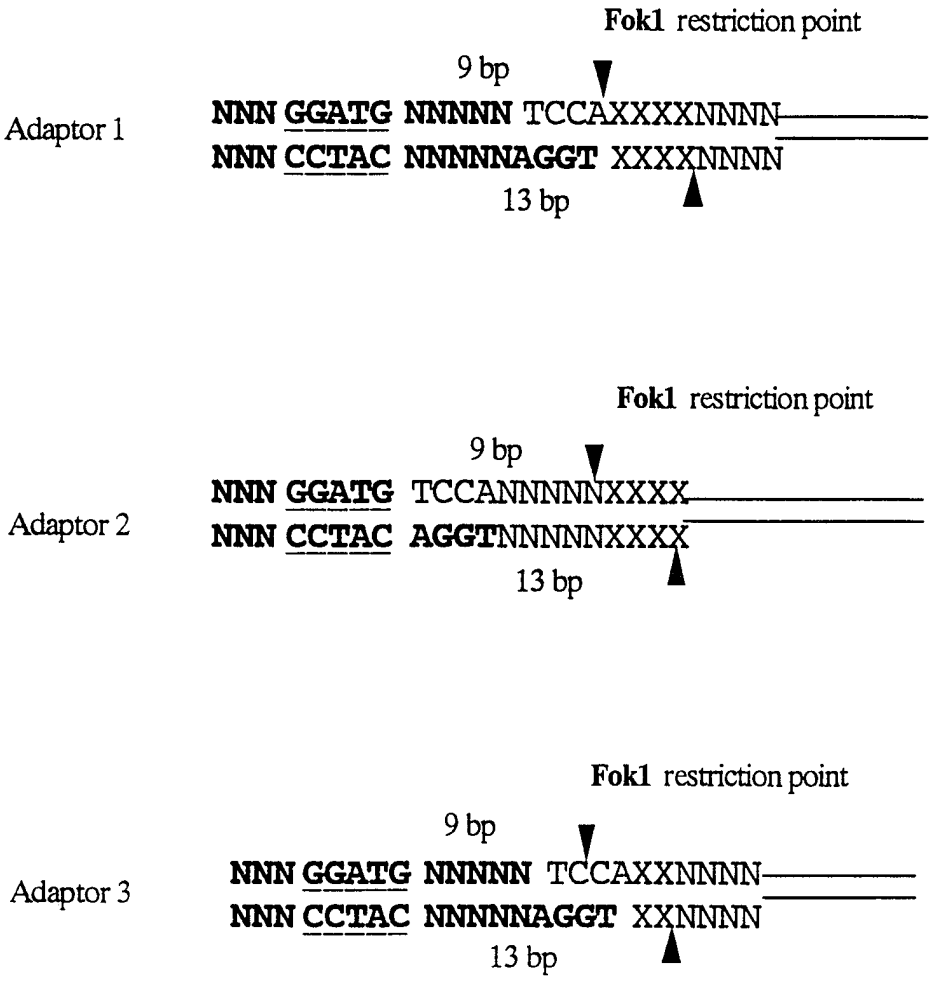

FokI restriction point
9 bp
Adaptor 1
NNN GGATG NNNNN TCCAXXXXNNNN
NNN CCTAC NNNNNAGGT XXXXNNNN
13 bp
FokI restriction point
9 bp
Adaptor 2
NNN GGATG TCCANNNNNXXXX
NNN CCTAC AGGTNNNNNXXXX
13 bp
FokI restriction point
9 bp
Adaptor 3
NNN GGATG NNNNN TCCAXXNNNN
NNN CCTAC NNNNNAGGT XXNNNN
13 bp

FIGURE 7

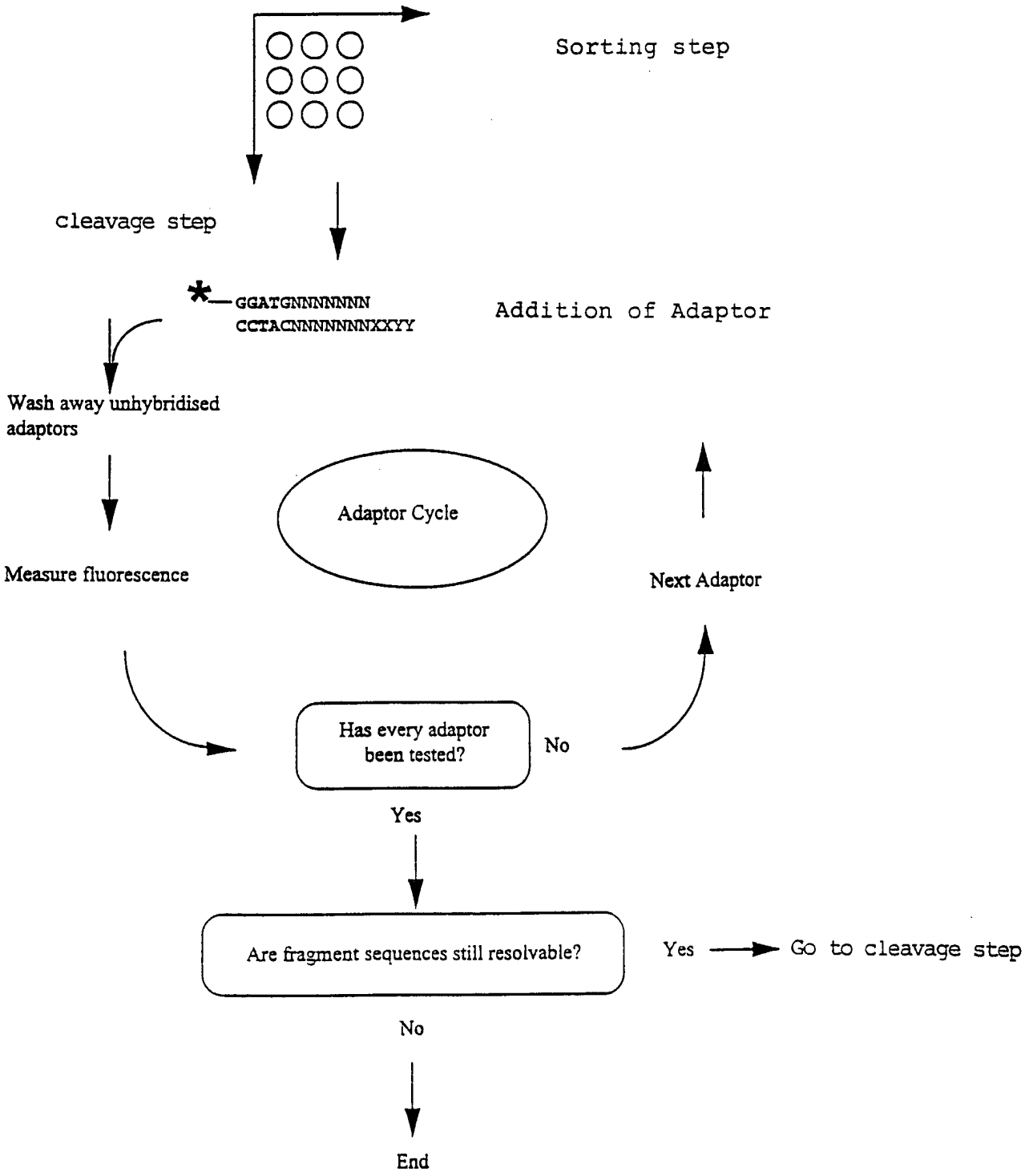
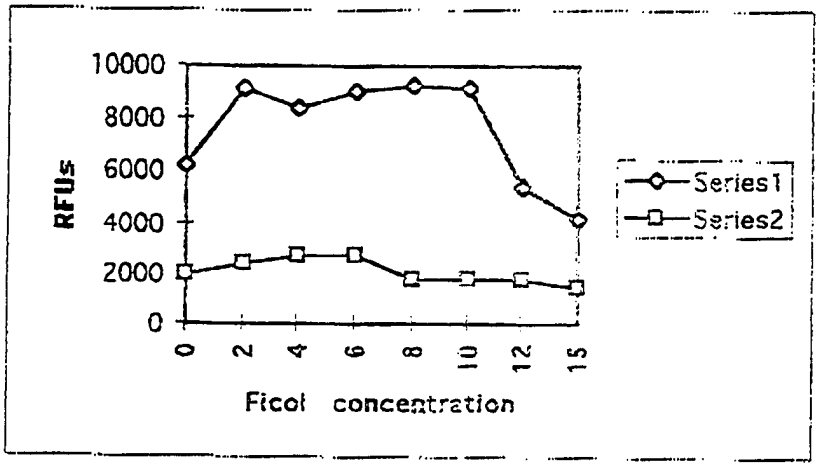
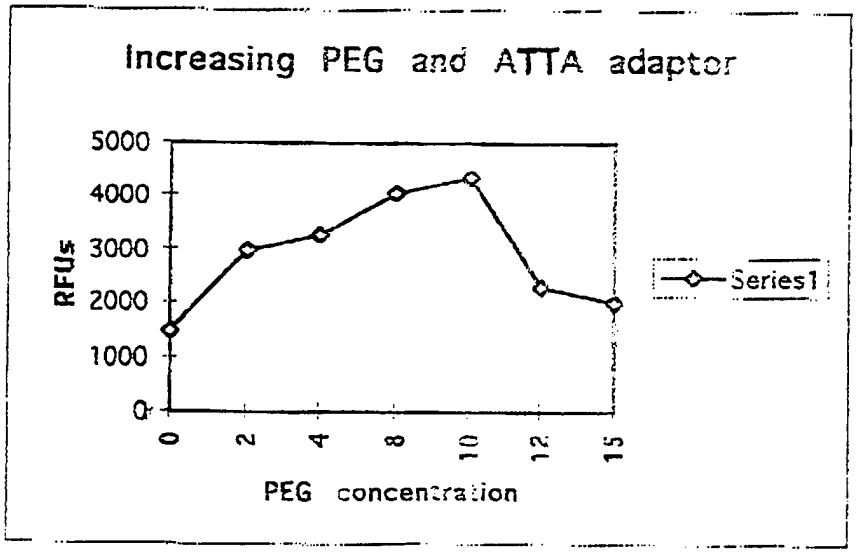

Sorting step
cleavage step
GGATGNNNNNNN
CCTACNNNNNNNXXYY
Addition of Adaptor
Wash away unhybridised adaptors
Adaptor Cycle
Measure fluorescence
Next Adaptor
Has every adaptor been tested?
No
Yes
Are fragment sequences still resolvable?
Yes
Go to cleavage step
No
End

FIGURE 8


10000
8000
6000
4000
2000
0
RFUs
0
2
4
6
8
10
12
15
Ficol concentration
Series1
Series2


Increasing PEG and ATTA adaptor
5000
4000
3000
2000
1000
0
RFUs
0
2
4
8
10
12
15
PEG concentration
Series1

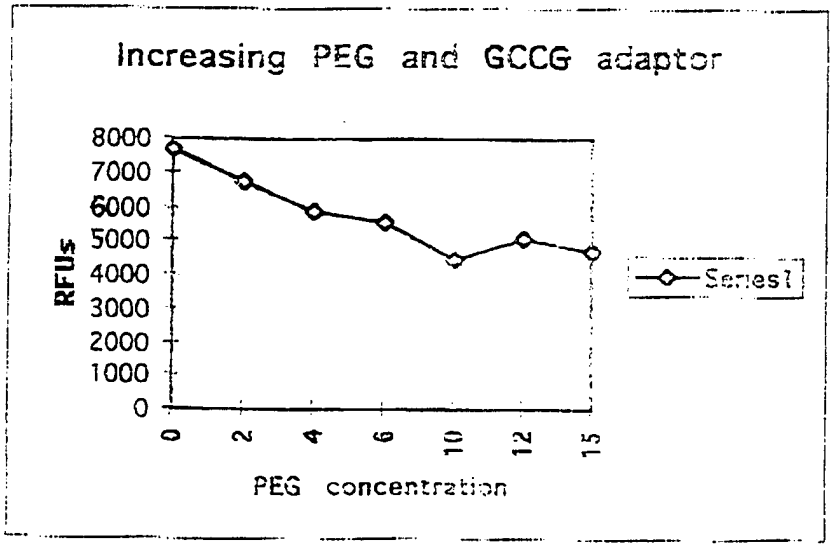

Increasing PEG and GCCG adaptor
8000
7000
6000
5000
4000
3000
2000
1000
0
RFUs
0
2
4
6
10
12
15
PEG concentration
Series1

FIGURE 9

# INTERNATIONAL SEARCH REPORT

Internat l Application No
PCT/GB 97/02734

A. CLASSIFICATION OF SUBJECT MATTER
IPC 6 C12Q1/68

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 6 C12Q

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

| Category ° | Citation of document, with indication, where appropriate, of the relevant passages | Relevant to claim No. |
|---|---|---|
| E | WO 97 46704 A (LYNX THERAPEUTICS INC) 11 December 1997<br>see the whole document | 1,7-13 |
| Y | WO 95 27080 A (LYNX THERAPEUTICS INC) 12 October 1995<br>see the whole document | 1-13 |
| Y | WO 95 20053 A (MEDICAL RES COUNCIL ;SIBSON DAVID ROSS (GB)) 27 July 1995<br>see the whole document | 1-13 |
| Y | WO 96 12014 A (LYNX THERAPEUTICS INC) 25 April 1996<br>see the whole document | 1-13 |

-/--

[X] Further documents are listed in the continuation of box C.

[X] Patent family members are listed in annex.

° Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

2

| Date of the actual completion of the international search | Date of mailing of the international search report |
|---|---|
| 24 February 1998 | 10/03/1998 |

| Name and mailing address of the ISA | Authorized officer |
|---|---|
| European Patent Office, P.B. 5818 Patentlaan 2<br>NL - 2280 HV Rijswijk<br>Tel. (+31-70) 340-2040, Tx. 31 651 epo nl,<br>Fax: (+31-70) 340-3016 | Osborne, H |

Form PCT/ISA/210 (second sheet) (July 1992)

# INTERNATIONAL SEARCH REPORT

Interna' l Application No
PCT/GB 97/02734

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

| Category | Citation of document, with indication,where appropriate, of the relevant passages | Relevant to claim No. |
|---|---|---|
| Y | WO 96 12039 A (LYNX THERAPEUTICS INC) 25 April 1996<br>see the whole document | 1-13 |
| Y | TANG K ET AL: "MATRIX-ASSISTED LASER DESORPTION/IONIZATION OF RESTRICTION ENZYME-DIGESTED DNA"<br>RAPID COMMUNICATIONS IN MASS SPECTROMETRY,<br>vol. 8, no. 2, February 1994,<br>pages 183-186, XP000608266<br>see the whole document | 1-13 |
| A | EP 0 309 969 A (DU PONT) 5 April 1989<br>see the whole document | 1 |
| A | UNRAU P ET AL: "Non-cloning amplification of specific DNA fragments from whole genomic digests using DNA 'indexers'"<br>GENE,<br>vol. 145, 1994,<br>pages 163-169, XP002056703<br>see the whole document | 1 |

2

# INTERNATIONAL SEARCH REPORT

Information on patent family members

Internal Application No
PCT/GB 97/02734

| Patent document cited in search report | Publication date | Patent family member(s) | Publication date |
|---|---|---|---|
| WO 9746704 A | 11-12-97 | NONE | |
| WO 9527080 A | 12-10-95 | US 5552278 A | 03-09-96 |
| | | AU 685628 B | 22-01-98 |
| | | AU 2379195 A | 23-10-95 |
| | | CA 2163662 A | 12-10-95 |
| | | EP 0703991 A | 03-04-96 |
| | | JP 8511174 T | 26-11-96 |
| | | US 5599675 A | 04-02-97 |
| | | US 5714330 A | 03-02-98 |
| WO 9520053 A | 27-07-95 | AU 1459595 A | 08-08-95 |
| | | EP 0739422 A | 30-10-96 |
| | | JP 9508268 T | 26-08-97 |
| WO 9612014 A | 25-04-96 | US 5604097 A | 18-02-97 |
| | | AU 3946195 A | 06-05-96 |
| | | AU 4277896 A | 06-05-96 |
| | | CZ 9700866 A | 17-09-97 |
| | | EP 0786014 A | 30-07-97 |
| | | EP 0793718 A | 10-09-97 |
| | | FI 971473 A | 04-06-97 |
| | | NO 971644 A | 02-06-97 |
| | | WO 9612039 A | 25-04-96 |
| | | US 5695934 A | 09-12-97 |
| | | US 5635400 A | 03-06-97 |
| | | US 5654413 A | 05-08-97 |
| WO 9612039 A | 25-04-96 | US 5695934 A | 09-12-97 |
| | | AU 3946195 A | 06-05-96 |
| | | AU 4277896 A | 06-05-96 |
| | | CZ 9700866 A | 17-09-97 |
| | | EP 0786014 A | 30-07-97 |
| | | EP 0793718 A | 10-09-97 |
| | | FI 971473 A | 04-06-97 |
| | | NO 971644 A | 02-06-97 |
| | | WO 9612014 A | 25-04-96 |
| | | US 5604097 A | 18-02-97 |
| | | US 5635400 A | 03-06-97 |
| | | US 5654413 A | 05-08-97 |

# INTERNATIONAL SEARCH REPORT

...ormation on patent family members

Internat... al Application No
PCT/GB 97/02734

| Patent document cited in search report | Publication date | Patent family member(s) | Publication date |
|---|---|---|---|
| EP 0309969 A | 05-04-89 | US 5102785 A | 07-04-92 |
| | | CA 1322942 A | 12-10-93 |
| | | DE 3854176 D | 24-08-95 |
| | | DE 3854176 T | 15-02-96 |
| | | DK 536388 A | 29-03-89 |
| | | IE 69577 B | 02-10-96 |
| | | JP 1137983 A | 30-05-89 |
| | | JP 2653684 B | 17-09-97 |