J. theor. Biol. (2003) **221,** 615–624
doi:10.1006/jtbi.2003.3211, available online at http://www.idealibrary.com on IDEAL®

Measurement of the Number of Molecules of a Single mRNA Species in a Complex mRNA Preparation

HUBERT HUG† AND RAINER SCHULER*‡

†*University Children's Hospital, Prittwitzstrasse* 43, *D*–89075 *Ulm, Germany and* ‡*Department of Theoretical Computer Science, University of Ulm, D*–89069 *Ulm, Germany*

(*Received on* 20 *February* 2002, *Accepted in revised form on* 19 *November* 2002)

The normalization of data obtained from hybridization experiments with DNA chips to determine mRNA expression and concentration (gene expression profiling) is an unsolved problem. Furthermore, slight changes in mRNA expression or small numbers of mRNA molecules which may be relevant to disease cannot be detected so far.

We have designed a method to calculate the number of molecules of a single mRNA species in a complex mRNA preparation. The basic concept is the transformation of a quantitative problem into a qualitative problem. Individual molecules pertaining to the same molecular species (IMPSMS) are transformed to a mixture of new different molecular species (DMS) and amplified.

We propose two implementations of the method. The first procedure is based on a method for cloning tagged nucleic acid molecules onto the surface of micro–beads. It should be possible to transform and determine up to 10^6 IMPSMS into new DMS. The second strategy uses multimeric linkers, a method frequently used in DNA computing to assemble random DNA. The second strategy should be easier to implement but is limited to a few hundred IMPSMS.

© 2003 Elsevier Science Ltd. All rights reserved.

1. Introduction

Growth and differentiation of a cell in a tissue is dependent on the expression levels of the genes. While the number and type of genes is usually identical in all cells of an organism, the mRNAs expressed and mRNA expression levels are specific for the cell type and state of differentiation.

RNA is not very stable and therefore, it is transcribed into complementary DNA (cDNA) for analysis. The cDNA concentrations of a large number of genes can be measured in parallel by using an addressable DNA micro–array or DNA chip (chips, 2002). Thereby the normalization of data obtained from different experiments is a difficult problem.

Currently it is not possible to label and count single mRNA molecules of a molecular species in a complex mixture. The most sensitive fluorescence–based assays for quantitation (e.g. PicoGreen) of nucleic acids in solution still need approximately 25 pg ml^{-1} of double–stranded DNA. Since quantitation of nucleic acids needs a sufficiently large quantity of molecules, an amplification step using PCR is necessary. Each round of a PCR amplification roughly doubles the number of DNA molecules

*Corresponding author. Tel.: +49-731-5024107; fax: +49-731-502412.
E-mail address: hubert.hug@medizin.uni–ulm.de (H. Hug), schuler@informatik.uni-ulm.de (R. Schuler).

0022-5193/03/$35.00 © 2003 Elsevier Science Ltd. All rights reserved.

PARSE EX1007.001

making a following quantitative analysis difficult.

In this paper, we make a theoretical suggestion on how to count the number of *individual molecules pertaining to the same molecular species* (IMPSMS) in a complex molecular sample. We use the fact that mRNA molecules with different sequences or of different length can be distinguished using standard methods. In other words, using standard biochemical methods it is possible to determine the number of *different molecular species* (DMS) in a sample. The method we propose relies on the transformation of the quantitative problem (count the number of IMPSMS) into a qualitative problem (detect new DMS). After transformation, the number of new DMS directly corresponds to the number of IMPSMS in the original sample.

We suggest two approaches to generate new DMS from IMPSMS.

- Targeted random labeling of each IMPSMS with an oligonucleotide tag.
- Targeted random addition of a variable number of short oligonucleotide linkers.

The first approach is based on a method for cloning tagged nucleic acid molecules onto the surface of micro–beads (Brenner *et al.*, 2000b). A tag consists of eight 4mers (blocks). To minimize cross–hybridization between tags and between mRNA molecules and tags, the number of 4mers used is restricted to eight (out of 256). During construction of the tags, one of the eight 4mers is selected for each block. The number of different tags that can be built in this way is approximately 1.6×10^7 (Brenner *et al.*, 2000b).

We suggest to randomly attach a tag to each IMPSMS in order to generate a new DMS. If the number of tags is sufficiently large, a unique tag will be attached to each IMPSMS and the number of new DMS is equivalent to the number of IMPSMS in the original sample. Each new DMS is uniquely defined by its tag sequence. In a final step the number of different tag sequences is determined. We consider two methods depending on the expected number of IMPSMS. If the number of required tag sequences is less than 60 000, the new DMS molecules are hybridized onto the surface of beads for further DNA–sequencing, using the massively parallel signature sequencing (MPSS) method (Brenner *et al.*, 2000a). For still smaller numbers, the sequencing can be avoided and the molecules are hybridized onto an addressable chip surface containing molecules with reverse complementary tag sequences. If the number of required tag sequences exceeds 60 000, the new DMS could be analysed by defined block selection as introduced below. Using this method in principle up to 10^6 IMPSMS can be detected.

The second approach to transform IMPSMS into new DMS is based on linker ligation. Short oligonucleotide linkers are mixed with a certain ratio of oligonucleotide terminators and are added to the IMPSMS preparation. At each IMPSMS a random number of oligonucleotide linkers will be added until elongation is terminated by adding an oligonucleotide terminator. New DMS are defined by the number of linkers attached to each IMPSMS. The length of the originating DMS can be determined (by gel electrophoresis), where we expect to be able to distinguish up to 1000 different linker lengths.

2. The Method

Nucleic acid molecules with different sequences or of different length can be distinguished using standard biochemical methods (Ausubel *et al.*, 2001; Sambrook *et al.*, 1989). This observation suggests the following approach for counting the number of molecules of a single molecular species:

1. Isolate the IMPSMS of the species of interest.
2. Transform IMPSMS into new DMS.
3. Amplify the new DMS.
4. Detect and count or estimate new DMS.

That is, we reduce a quantitative problem "estimate the number of IMPSMS" into a qualitative problem "detect new DMS".

The first and the third step rely on established methods in molecular biology (see below). We propose two methods to transform the IMPSMS into new DMS. Both methods use the fact the hybridization of DNA molecules with identical sequences is competitive and therefore, the tags are selected randomly. In both cases, it is

PARSE EX1007.002

possible to distinguish the new DMS by either gel electrophoresis, or labeling and separation, or DNA-sequencing methods.

Nucleic acid molecules can differ in length and in the sequence of the base pairs. In order to transform IMPSMS into new DMS, oligonucleotides with different sequences (tag labeling) or of different length (multimeric linker labeling) are attached to the IMPSMS. In Sections 2.1 and 2.2 we describe the tag labeling and multimeric linker labeling methods, respectively. The implementation of the methods is described in Sections 2.3 and 2.4.

2.1. TAG LABELING

Oligonucleotides of the same length but with different sequences of base pairs (tags) are used to label the IMPSMS. The tags that are attached to the IMPSMS are selected randomly and independently for each IMPSMS. If the number of tag sequences is sufficiently large, it is very likely that almost every IMPSMS will be combined with a unique tag sequence.

To estimate the expected number of identical tags, let m denote the total number of different tag sequences, and let l denote the (unknown) number of IMPSMS. In the following calculation we assume that the tag sequences are numbered 1 to m and that the IMPSMS are attached to tags with sequences $i_1, i_2, \ldots, i_l$ in this order.

The probability that the n–th tag sequence is attached to the j–th IMPSMS is $1/m$ and the probability that the n–th tag sequence is selected by some IMPSMS is $1-(1-1/m)^l$. Let X_n be an indicator variable for this event, i.e.

$$X_n = \begin{cases} 1 & \text{if } n \in \{i_1, i_2, \ldots, i_l\}, \\ 0 & \text{otherwise.} \end{cases}$$

Then $E[X_n] = 1-(1-1/m)^l$ is the expected value of X_n and the expected value of the number of different tag sequences $X = \sum_{n=1}^{m} X_n$ is given by (using the fact that $(1-1/m)^m \approx 1/\mathrm{e}$ for sufficiently large m),

$$E[X] = E[\sum_{n=1}^{m} X_n] = \sum_{n=1}^{m} E[X_n] \approx m(1-\mathrm{e}^{-1/m}).$$

For example, if $m = 10^7$, the expected number of identical tag sequences is 5 for 10^4 IMPSMS, and increases to approximately 500 and 50 000 for 10^5 and 10^6 IMPSMS, respectively.

2.2. MULTIMERIC LINKER LABELING

In this method, multimeric linkers of different lengths are attached to the IMPSMS. Each multimeric linker consists of a certain number of oligonucleotide linkers (repeats) followed by an oligonucleotide terminator. New DMS are defined by the number of repeats of the multimeric linkers attached to the IMPSMS.

During the transformation as described in Section 2.4, Cases 1–3, linker oligonucleotides are added to the IMPSMS (by hybridization) until the hybridization is stopped by a terminator oligonucleotide. The probability that a terminator is selected depends of the ratio p of oligonucleotide terminators in the oligonucleotide linker preparation. The probability, that a sequence of oligonucleotide linkers is terminated, is given by p, and the probability that a further linker is added, is given by $q = (1-p)$.

The number of repeats is (ideally) described by a geometric distribution, in this case the probability of exactly n repeats is $p\ q^n$. To estimate the expected number of different lengths let l denote the (unknown) number of IMPSMS and let m be some upper bound on the number of repeats possible in a multimeric linker. In the following calculation we assume that the IMPSMS are numbered 1 to l and that the multimeric linkers attached to the IMPSMS have lengths $i_1, i_2, \ldots, i_l$ in this order. The probability that a multimeric linker of length n is attached to some IMPSMS is $1-(1-p \cdot q^n)^l$. Let X be an indicator for this event, i.e.

$$X_n = \begin{cases} 1 & \text{if } n \in \{i_1, i_2, \ldots, i_l\}, \\ 0 & \text{otherwise.} \end{cases}$$

Then $E[X_n] = 1-(1-p \cdot q^n)^l$ is the expected value of X_n and the expected value of the number of different multimeric linker length

PARSE EX1007.003

$X = \sum_{n=1}^{m} X_n$ is given by (using the fact that $1 - x \leqslant e^{-x}$ for $x \leqslant 1$),

$$E[X] = E[\sum_{n=1}^{m} X_n] = \sum_{n}^{m} E[X_n] \geqslant \sum_{n=1}^{m} (1 - e^{-lpq^n}).$$

Consider a ratio $p = 1/300$ (expected number of repeats $1/p = 300$) and multimeric linker of length up to $m = 1200$. Then, using the above inequality, the expected number of identical repeats is less than 10 for 100 IMPSMS (10%) and less than 34 for 200 IMPSMS (17%).

We observe that the number of repeats is approximately geometrically distributed (depending on the ratio of oligonucleotide terminators) making short multimeric linkers more likely. In contrast to tag labeling (and Case 4 below), where we assume a uniform distribution, it is therefore more likely that identical linkers will be attached to more than one IMPSMS.

During the transformation as described in Section 2.4 (Case 4) multimeric linkers of different lengths are prepared and then attached to the IMPSMS. As with tag labeling, we assume that the multimeric linkers which are added to the IMPSMS are chosen randomly and independently for each IMPSMS. In this case the probability that the linker attached to an IMPSMS has length n is $1/m$, where m is the number of different multimeric linker lengths.

The number of different multimeric linker length attached to the IMPSMS are uniformly distributed and can be calculated as for tag labeling. The expected number of different multimeric linker length is approximately $m(1 - e^{-l/m})$, where l is the number of IMPSMS. For example, if $m = 1200$ then the expected number of identical multimeric linker length is 5 for 100 IMPSMS and increases to 16 for 200 IMPSMS.

To increase the applicability of multimeric linker labeling the number of IMPSMS can be estimated e.g. by a maximum likelihood estimation, from the number of different multimeric linker lengths, or the maximal multimeric linker length detected (Hollas & Schuler, 2002).

2.3. IMPLEMENTATION OF TAG LABELING

The oligonucleotides containing the tag sequences, e.g. 32–mers, can be designed and constructed as described (Brenner *et al.*, 2000b). A tag is an oligonucleotide and consists of eight 4mers (blocks). In total there are $4^4 = 256$ different 4mers from which eight are used to build all the tags. That is, each block of the tags consists of one of the eight selected 4mers. The number of different tags that can be constructed in this way is approximately 1.6×10^7 (Brenner *et al.*, 2000b). Furthermore, the different tags are present in the same quantity.

The tags are used to separate DNA molecules. Therefore, in a solution containing (some) tags and oligonucleotides with reverse complementary tag sequences, each tag should bind only to the oligonucleotide with the reverse complementary tag sequence. To minimize cross–hybridization, the number of different 4mers of the tags is restricted to eight (Brenner *et al.*, 2000b). The word design e.g. avoids palindromic 4mers and 4mers with partially identical base pairs sequences.

We now give a description of the individual steps involved in the tag labeling method. To separate the mRNA species of interest from the mRNA pool we use oligonucleotides consisting of the reverse complementary sequence of a part of the mRNA sequence followed by a random tag sequence, a restriction enzyme recognition site (R) and an unique primer binding site (abbreviated as PBS in Fig. 1). The 5′ end of the primer binding site is labeled with biotin in order to conjugate paramagnetic streptavidine beads (Bio/Str) to separate the mRNA species from the mRNA pool with a magnet (Fig. 1).

Now the first strand of a cDNA is synthesized (Fig. 1). Note that each cDNA contains a tag sequence. Furthermore, with very high probability, the number of different tag sequences will be very close to the number of mRNA molecules of the species under consideration.

Depending on a suitable internal forward primer binding site, a part of the cDNA sequence is amplified by the polymerase chain reaction (PCR). At the 5′ ends of the tags, a unique sequence which is used as binding site

PARSE EX1007.004

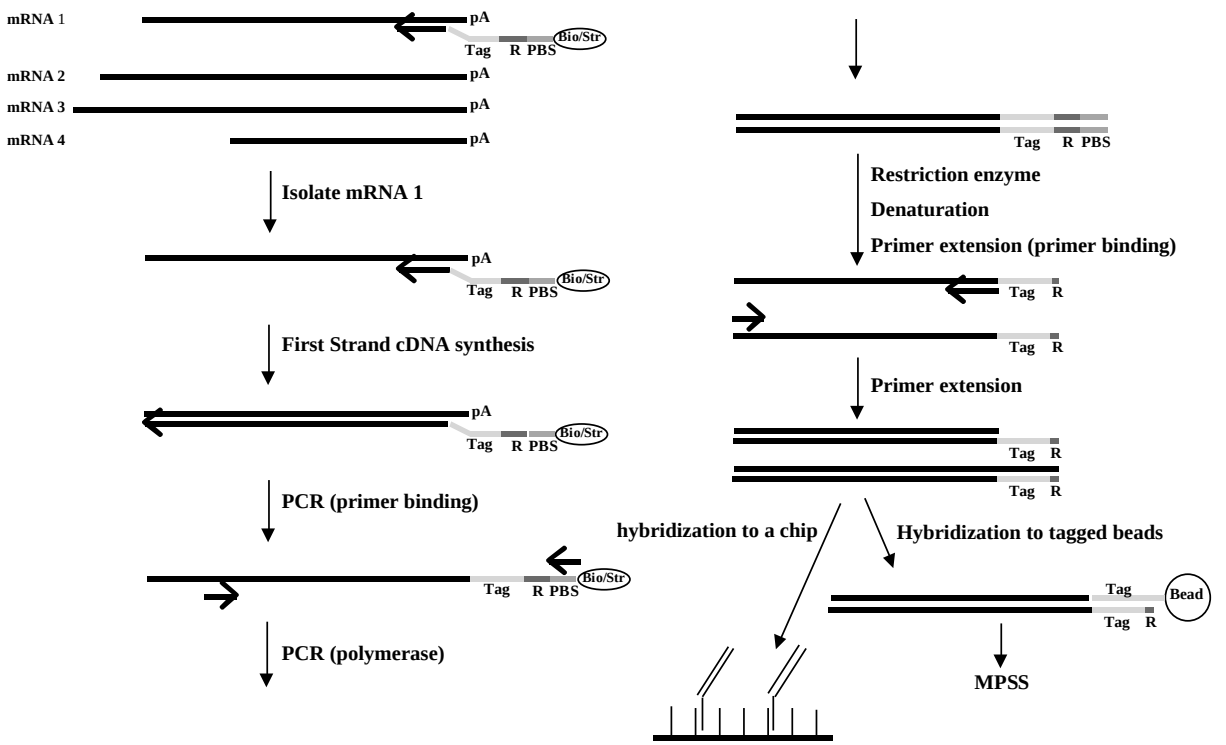


FIG. 1. Overview for the determination of the absolute number of mRNA molecules of a sequence of interest (IMPSMS) in a complex mRNA mixture by tagging.The molecules of mRNA1 are converted to new DMS. Only one tag sequence is shown. The steps shown are: hybridization of tag sequences to the total RNA or mRNA preparation, isolation of new DMS via paramagnetic streptavidine, first strand cDNA synthesis with reverse transcriptase, amplification by PCR, removal of PBS by restriction cutting, primer extension to with DNAse free Klenow polymerase to create double–stranded DNA leaving a single–stranded tag, (1) hybridization to tagged beads for further sequencing by MPSS or (2) direct hybridization to a DNA chip containing reverse complementary tag sequences. pA, polyA tail; Bio/Str, biotin/paramagnetic streptavidine; PBS, primer binding site; R, restriction endonuclease recognition site.

for the reverse PCR primers, is located (PBS in Fig. 1). After PCR amplification, the primer binding site sequence is removed by cutting with a restriction enzyme between the tag and the primer binding site sequences (R) (Fig. 1).

The PCR products can be hybridized to tagged beads (Brenner *et al.*, 2000 a, b; Tyagi, 2000), where the sequences attached to the beads are reverse complementary to the tag sequences. After hybridization, the beads loaded with DNA are separated as described (Brenner *et al.*, 2000b). Then the number of different tag sequences is determined by MPSS (Brenner *et al.*, 2000a). The MPSS method is able to sequence 20 bp of approximately 60 000 sequences attached to beads, respectively. To bind the adaptors for MPSS, the cDNAs are cut at an internal unique restriction enzyme recognition sequence near the tag sequences.

Currently, the sequencing of a large number of tagged DNAs is only feasible in laboratories with the sophisticated equipment for MPSS (e.g. Lynx Therapeutics). Therefore, an alternative strategy to determine the number of different tag sequences could be the use of DNA chips, which consist of oligonucleotides of the reverse complementary tag sequences. The maximum number of oligonucleotides that can be attached to a chip surface is approximately 50 000 (chips, 2002). The target for chip hybridization is synthesized by primer extension (Fig. 1). The target is labeled with a fluorescence dye. The number of the original mRNA molecules corresponds to the signals generated on the chip surface.

The limit of 20 bp which can be sequenced using the MPSS method restricts the number of tags which can be distinguished to $8^5 = 32\,768$.

PARSE EX1007.005

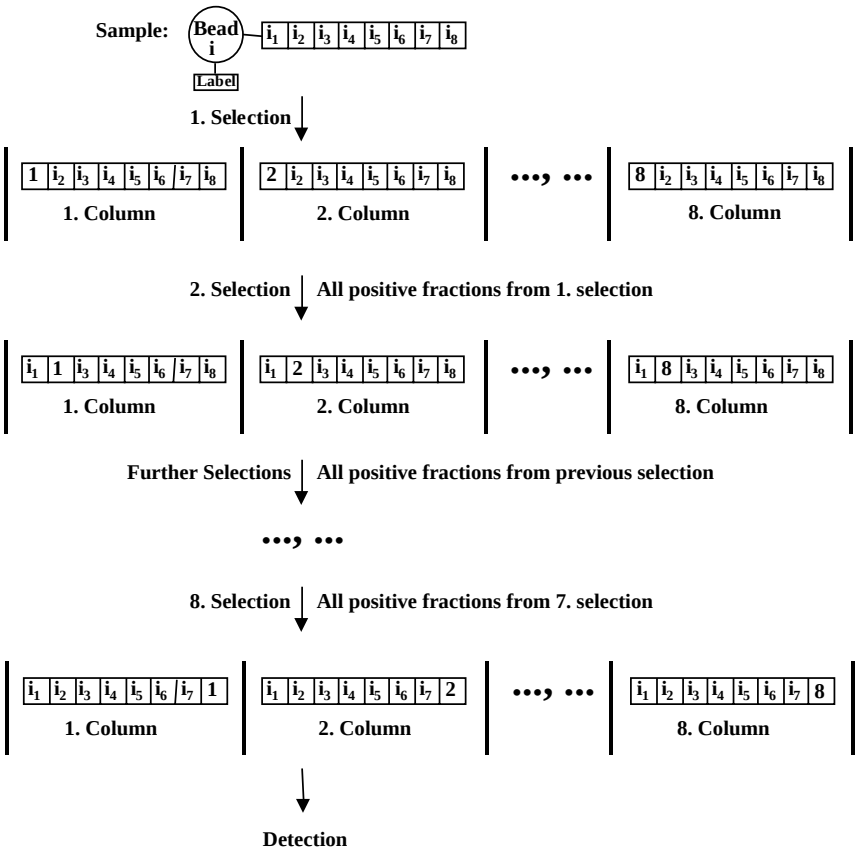


FIG. 2. Defined block selection. Attached to the beads are the tags consisting of eight blocks, denoted by *i*1–*i*8. A block *i* consists of one out of eight chosen 4mers. Hybridized beads *i* are recovered from the primer extension reaction (Fig. 1). Beads are separated through eight rounds of selection. In the first selection the beads are sequentially run over the eight columns with a defined 4mer at block position 1. The matrix for the oligonucleotides in the columns is a polysaccharide (e.g. sepharose). The temperature is selected in order that 100% matches are captured but beads containing oligonucleotides with one or more mismatches are in the flow through. Elution is above the melting temperature of the oligonucleotides. The following selection rounds are only done for positive samples of the previous selection round. The beads contain either a fluorescence or a magnetic label to allow detection.

To extend the method, we suggest defined block selection (Fig. 2). For each block position of the tag we prepare eight columns. In each column, the 4mer of one block is defined (Fig. 2). All tags which contain the reverse complementary 4mer at this position are retained and detected. The detection is either through fluorescently labeled or magnetic beads. The tagged beads of the positive columns are eluted and applied to the next set of columns. In this set of columns the next block consists of defined 4mers (Fig. 2). Using this strategy we can use the whole set of $8^8 \approx 1.6 * 10^7$ tags. Estimating the diameter of a bead to 2 μm, the required volume of the column would be $\geqslant 0.13$ μl. Instead of columns the oligonucleotides could be covalently attached to poly–acrylamide gels (Braich *et al.*, 2001, 2002). Separation could also be done in micro–flow reactors (McCaskill, 2001).

2.4. IMPLEMENTATION OF MULTIMERIC LINKER LABELING

The second approach uses the fact that the length of DNA molecules can be determined. In fact, measuring the length of DNA molecules by gel electrophoresis is the basis for standard DNA sequencing. Using available kits, it is possible to distinguish approximately 1000 different lengths in one run. Therefore, this approach should be easier to implement but is limited to determine the number of a few hundred mRNA molecules.

PARSE EX1007.006

We observe that the multimeric linker labeling method codes the number of IMPSMS in unary, i.e. in the number of different lengths of the multimeric linker.In contrast, the tag labeling method described in the previous section uses a four letter alphabet to count the number of IMPSMS.

The self-assembly of DNA is used to build multimeric linkers of different length. Consider oligonucleotides P and Q such that the sequence of P at the 5′ end is reverse complementary to the sequence of Q at the 3′ end and the sequence of P at the 3′ end is reverse complementary to the sequence of Q at the 5′ end. [For example $P = \text{CTAGAACC}$ and $Q = \text{CTAGGGTT}$ see Fig. 3(B) and (C).] If P and Q are put into solution, oligonucleotides P and Q will hybridize (at the 5′ and 3′ ends) and build chains of P's and Q's of different length. It is interesting to note, that self-assembly of DNA is used in DNA computing to generate a set of all possible combinations of different oligonucleotide sequences (see e.g. Adleman, 1994). Here, we use identical oligonucleotides for multimerization.

We now give a description of the individual steps involved in the multimeric linker labeling method. The first strand of the cDNA is synthesized by using an oligonucleotide whose sequence is reverse complementary to the sequence of the mRNA of interest. At the 5′end of this oligonucleotide a recognition site of a restriction endonuclease (e.g. BamH1) is attached [Fig. 3(A)]. A double stranded cDNA is synthesized and cleaved with the restriction endonuclease (e.g. BamH1) according to standard protocols (Ausubel *et al.*, 2001) as outlined in Fig. 3(A).

Linkers consist of the recognition site of BamH1 and a suitable short sequence which minimizes cross–hybridization [Fig. 2(B)]. Linkers are mixed with a certain ratio of stop sequences [Fig. 2(B)] while being added to the cDNAs in order to hybridize to the sticky BamH1 ends (Fig. 2C). The stop linker will terminate linker ligation at every possible position in analogy to the dideoxy–method of DNA sequence analysis (Sanger *et al.*, 1977). Again we assume that the linker ligation is a random process.

In the final step the multimeric linkers are ligated with T4 DNA ligase to the cDNA and with each other. Each cDNA molecule should now have attached a multimeric linker of a particular length [Fig. 2(C)]. The cDNAs are amplified by PCR and the number of different multimeric linkers is determined by DNA length analysis using gel electrophoresis with appropriate DNA size markers.

In practice, linker ligation can be carried out following several alternative protocols:

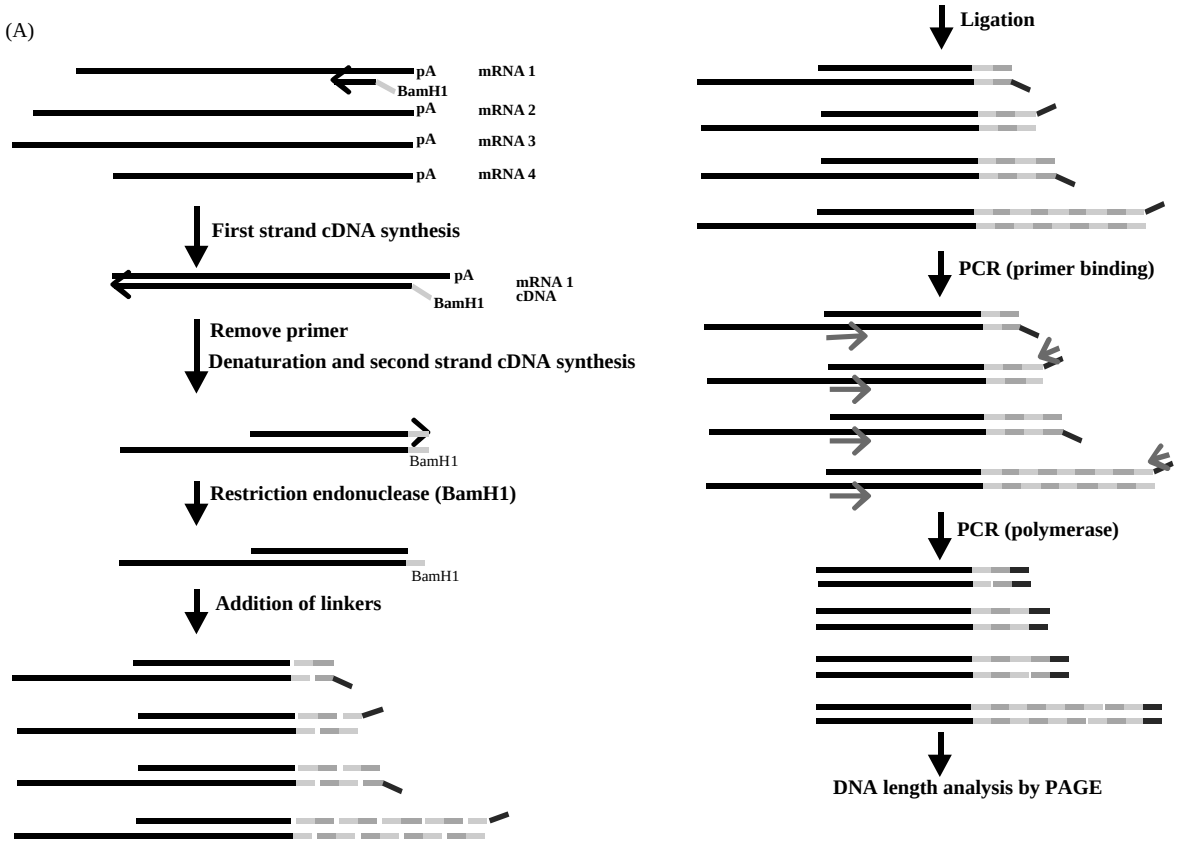
1. The double stranded cDNA with sticky ends is mixed together with linker and terminator oligonucleotides. Linker and terminator oligonucleotides are at a defined ratio.
2. First the linker oligonucleotides are added and ligated to the cDNA preparation. Terminator oligonucleotides are added and ligated after a defined time interval.
3. First linker and terminator oligonucleotides are mixed at a defined ratio and ligated. The products are then added to the cDNA preparation and ligated.
4. A similar strategy is to chemically synthesize DNA oligonucleotides of different lengths and hybridize them. One strand will yield a BamH1 sticky end. The advantage would be that the length of the oligonucleotides can differ in only 1 bp. The synthetic molecules are then added to the cDNA preparation and ligated.

We observe that the length of a new DMS, i.e. the number of linker oligonucleotides of a multimeric linker molecule, is geometrically distributed in Cases 1–3, and uniformly distributed in Case 4. The probability of the geometric distribution is determined by the ratio of linker and terminator oligonucleotides.

3. Discussion

The methods described have an immediate application for normalizing target RNAs for DNA chip hybridization. The advantage is the detection of IMPSMS by converting them into DMS. Prerequisites of the methods are that the conditions for the critical steps like the annealing of the tag sequences to the mRNA molecules, first–strand cDNA synthesis and the ligation steps are close to 100% efficiency. Mispairings

PARSE EX1007.007



(B)

Linker sequences:
CTAGAACC, CTAGGGTT

Terminator Sequences with PCR primer binding sites:
TTGGTCGGATCGGTATATCCGA
TCGGATATACCGATCCGAGATC

PCR primers:
5′Primer: binding to cDNA of interest
3′Primer: TCGGATATACCGATCCGA

(C)

CTAGAACC
GATCTTGGTCGGATCGGTATATCCGA
cDNA

CTAGAACCCTAGAGCCTAGCCATATACCCA
GATCTTGGGATC
cDNA

CTAGAACCCTAGAACC
GATCTTGGGATCTTGGTCGGATCGGTATATCCGA
cDNA

FIG. 3. Overview for the determination of the absolute number of mRNA molecules of a sequence of interest (IMPSMS) in a complex mRNA mixture by multimeric linker ligation. (A) The molecules of mRNA1 are converted to new DMS. The steps shown are: hybridization of a specific primer containing a BamH1 site to the IMPSMS, first and second strand cDNA synthesis with reverse transcriptase, cutting with BamH1 to create 5′–overhanging sticky ends, addition of linkers with a defined ratio of oligonucleotide terminators to create DMS of different lengths, ligation with T4 DNA ligase, amplification by PCR with the forward primer consisting of specific sequences of the cDNA under investigation and the reverse primer hybridizing to the non–hybridizing part of the terminator, DNA length determination by PAGE. pA, polyA tail; PAGE, poly–acrylamide gel electrophoresis. (B) Proposed sequences of the oligonucleotides to form multimeric linkers and the terminator sequences which include the primer binding site for PCR. In the terminator sequences, the sequences which do not allow further linker hybridization and which are the PCR 3′–primer binding site are underlined. (C) Formation of the multimeric linkers. Only the first three linker ligation products are shown.

of the tag sequences should be reduced by the selection of the proper 4mers. The recognition sequences for the restriction enzymes involved in the manipulation steps must not occur in the part of the selected cDNA sequence. If it is impossible to find such a region in the cDNA

PARSE EX1007.008

sequence other restriction enzymes have to be chosen.

After the initial conversion, the number of DMS is equal to the number of IMPSMS. The following manipulation steps are not quantitative. The methods should first be established with cloned cDNA sequences. The number of these cDNA molecules can be roughly estimated by staining with a fluorescent dye and absorption at 260 nm or gel electrophoresis with DNA size markers of known molarity. If the number of DMS is in the expected range, the methods should be established with mRNAs coding for proteins with basic functions and which are present in all eukaryotic cells, e.g. actin, tubulin, GAPDH, histones or other so-called housekeeping genes (Hsiao *et al.*, 2001). These mRNAs are present in relatively high amounts and are expressed with minimal fluctuations. When this is established, one could start to determine the number of tissue– or tumor–specific mRNAs, which usually show a lower expression. In some cases, the number of mRNA molecules could also be estimated by competitive PCR (Wang *et al.*, 1989; Gilliland *et al.*, 1990). Our protocol may also be established using such data.

In principle, our strategies provide a very accurate method for measuring the expression of defined candidate genes involved in diseases. Expression of these genes could be determined in parallel to large-scale gene expression profiling. For instance, low amounts of mRNAs expressed by viruses could be detected at earlier stages of infection.

In many diseases which could be diagnosed according to gene activity, i.e. the mRNA expression pattern, the problem is that only very little tissue amounts are available. In tumor surgery (e.g. brain tumors) it is often not possible to obtain enough control tissue. Our method could be suitable to detect changes in mRNA levels in total RNA or mRNA preparations which are derived of only a few cells. In addition, information about the exact number of IMPSMS is provided. In molecular diagnosis, the mRNAs under investigation have to be compared to the number of molecules of control mRNAs of housekeeping genes which are supposed not to be influenced by the disease.

A further application is to determine the number of a certain mRNA (IMPSMS) species in a single cell, when the number of cells is counted before RNA isolation. So far, it is only possible to determine relative mRNA concentrations in a total RNA or mRNA preparation with standard methods (e.g. Northern blotting, RNAse protection assay, quantitative RT–PCR, filter dot blots or chip hybridization). Recently, a combination of RT–PCR and flow cytometry has been used to show that endothelial cells contain approximately 400–900 copies of plasminogen activator inhibitor–1 mRNA copies per cell (Stemme *et al.*, 2001). Our method avoids the uncertainty introduced by PCR steps before the IMPSMS are transformed to new DMS and therefore, should yield more precise values.

Finally, our method can be extended to determine the exact number of molecules that are spotted onto a position of a DNA (or RNA) chip. The amount of nucleic acid spotted at a defined position of a chip is in the range of 200 fmol (XNA on gold, ThermoHybaid). The molecules bound to such a position could be tagged and then amplified by PCR in the presence of a fluorescently labeled primer. The double stranded PCR–products go into solution and after denaturation can be directly used for hybridization with oligonucleotide chips consisting of the reverse complementary tag sequences. Again, identical molecules (IMPSMS) are made distinguishable with different tags and the tags are then counted to yield the number new DMS.

We thank the anonymous referees for comments which helped to improve the paper.

REFERENCES

Adleman, L. (1994). Molecular computation of solutions to combinatorial problems. *Science* **266,** 1021–1024.

Ausubel, F., Brent, R., Kingston, R., Moore, D., Seidman, J., Smith, J. & Struhl, K., eds (2001). *Current Protocols in Molecular Biology*. New York: Wiley and Sons.

Braich, R., Johnson, C., Rothemund, P. W. K., Hwang, D., Chelyapof, N. & Adleman, L. (2001). Solution of a satisfiability problem on a gel–based DNA computer. In: DNA 2000, (Condon, A., ed.), vol. 2054, pp. 27–42. Berlin: Springer Verlag.

PARSE EX1007.009

Braich, R., Chelyapof, N., Johnson, C., Rothemund, P. & Adleman, L. (2002). Solution of a 20–variable 3–SAT problem on a DNA computer. *Science* **296,** 499–502.

Brenner, S., Johnson, M., Bridgham, J., Golda, G., Lloyd, D., Johnson, D., Luo, S., McCurdy, S., Foy, M., Ewan, M., Roth, R., George, D., Eletr, S., Albrecht, G., Vermaas, E., Williams, S., Moon, K., Burcham, T., Pallas, M., DuBridge, R., Kirchner, J., Fearon, K., Mao, J. & Corcoran, K. (2000a). Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat Biotechnol* **18,** 630–634.

Brenner, S., Williams, S., Vermaas, E., Storck, T., Moon, K., McCollum, C., Mao, J., Luo, S., Kirchner, J., Eletr, S., DuBridge, R., Burcham, T. & Albrecht, G. (2000b). In vitro cloning of complex mixtures of DNA on microbeads: physical separation of differentially expressed cDNA's. *Proc. Natl Acad. Sci. U.S.A.* **97,** 1665–1670.

Chips. (2002). http://www.gene-chips.com

Gilliland, G., Perrin, S., Blanchard, K. & Bunn, H. (1990). Analysis of cytokine mRNA and DNA: detection and quantitation by competitive polymerase chain reaction. *Proc. Natl. Acad. Sci. U.S.A.* **97,** 2725–2729.

Hollas, B. & Schuler, R. (2002). A stochastic approach to count RNA molecules using DNA sequencing methods. Submitted.

Hsiao, L.–L., Dangond, F., Yoshida, T., Hong, R., Jensen, R. V., Misra, J., Dillon, W., Lee, K. F., Clark, K. E., Haverty, P., Weng, Z., Mutter, G. L., Frosch, M. P., MacDonald, M. E., Milford, E. L., Crum, C. P., Bueno, R., Pratt, R. E., Mahadevappa, M., Warrington, J. A., Stephanopoulos, G., Stephanopoulos, G. & Gullans, S. R. (2001). A compendium of gene expression in normal human tissues. *Physiol. Genomics* **7,** 97–104.

McCaskill, J. S. (2001). Optically programming DNA computing in microflow reactors. *Biosystems*, **59,** 125–138.

Motwani, R. & Raghavan, P. (1995). *Randomized Algorithms*. Cambridge: Cambridge University Press.

Sambrook, J., Fritsch, E. & Maniatis, T. (1989). Molecular Cloning. *A Laboratory Manual*. 2nd edn, New York: Cold Spring Harbor Laboratory Press.

Sanger, F., Nicklen, S. & Coulson, A. (1977). DNA sequencing with chain–terminating inhibitors. *Proc. Nat. Acad. Sci. U.S.A.* **74,** 5463–5467.

Stemme, V., Rymo, L., Risberg, B. & Stemme, S. (2001). Quantitative analysis of specific mRNA species in minute cell samples by RT–PCR and flow cytometry. *J. Immunol. Methods* **249,** 223–233.

Tyagi, S. (2000). Taking a census of mRNA populations with microbeads. *Nat. Biotechnol.* **18,** 597–598.

Wang, A., Doyle, M. & Mark, D. (1989). Quantitation of mRNA by the polymerase chain reaction. *Proc. Natl. Acad. Sci. U.S.A.* **86,** 9717–9721.

PARSE EX1007.010