Nucleic Acids Research, 2001, Vol. 29, No. 22 e114

Preparing a human membrane and secreted protein-enriched cDNA library using PCR primers derived from a genomic database

Yi Fan[1], Chih Yuan Wu[2], Cheng Wei Chen[3], Tse Wen Chang[1,4] and Carmay Lim[2,5,*]

[1]Department of Life Science, [2]Department of Chemistry and [3]Department of Computer Sciences, National Tsing Hua University, Hsinchu 300, Taiwan, [4]Development Center for Biotechnology, Taipei, Taiwan and [5]Institute of Biomedical Sciences, Academia Sinica, Taipei 11529, Taiwan

Received June 28, 2001; Revised September 21, 2001; Accepted September 29, 2001

ABSTRACT

We describe here a strategy for preparing a human membrane and secreted protein (MSP)-enriched cDNA library based on human MSP- and non-MSP-encoding cDNA sequences in the databases. The signal peptide parts of the MSP-encoding cDNA sequences, which currently comprise about half of the estimated total number in humans, were analyzed for common patterns. These patterns form a 'minimal' set of polymerase chain reaction primer candidates of length varying from 9 to 21 nt. The products stemming from each primer candidate were determined and the results allowed us to obtain an 'optimal' mixed-length primer set. Ninety-six percent of the primers in this set were predicted to yield ≤10% undesired products, and the desired MSP-cDNA products could be easily separated by gel electrophoresis. The present analysis establishes a methodology for preparing a cDNA library that enables the analysis of individual MSPs. This methodology may also help identify new MSPs. As many cell regulatory processes are mediated by secreted proteins and their membrane-bound receptors, the preparation of a MSP-enriched cDNA library should benefit research on MSPs.

INTRODUCTION

Membrane and secreted proteins (MSPs) play important roles in mediating cell–cell interactions, cell growth and differentiation, activation and apoptosis (1–3). They include receptors, ion channels, cytokines, cell adhesion molecules, extracellular matrix proteins, hormones, immunoglobulins and plasma proteins. These proteins are characterized by an N-terminal signal peptide consisting of typically 15–30 amino acids, which directs their initial transportation through the membrane of endoplasmic reticulum before delivery to membrane enclosed organelles or plasma membrane or to be secreted. A signal peptide consists typically of three regions: (i) a usually positively charged, short N-region; (ii) a hydrophobic, 7–15 residue H-region; and (iii) a relatively polar, 3–7 residue C-region that contains the signal peptidase cleavage site (4,5).

Due to the biological importance of MSPs, several groups have developed methods to identify genes encoding these proteins. The signal-sequence trap method is based on the ability of cDNA sequences (cDNAs) encoding signal peptides to redirect the expression of an export defective receptor gene to the cell surface of mammalian cells (1,2,6,7). A rather ingenious signal trap is based on the ability of cDNAs encoding MSPs to provide a functional signal sequence to an N-terminal truncated invertase enzyme, whose secretion is essential for yeast cells to grow in a sucrose medium (1,2). The signal-exon trap combines features of signal sequence trap methods with the exon trap: exons are trapped and translated, which allows screening for signal peptide function (8). Another approach combines a cDNA library that is derived from rough endoplasmic reticulum-bound mRNA with a high throughput *in situ* hybridization procedure and subsequent sequence analysis (9). The DNA microarray method determines the relative abundance of individual transcripts bound to membrane and in the cytosol (10). However, none of the aforementioned methods have been used to construct a human MSP-enriched cDNA library due perhaps to various technical difficulties inherent in each method (see Discussion).

As MSPs are closely associated with regulatory or disease processes, a library of these proteins would be of scientific and application value. A library enriched in human MSP-coding cDNAs can be used to screen unknown receptor–ligand pairs for a given surface receptor or secreted protein. Such a library could save screening cost and time compared with using a library of all human cDNAs, as MSPs comprise only an estimated 10% (11) of the total number of genes (estimated to be ≈40 000) in the human genome (12). Another application of

*To whom correspondence should be addressed at: Institute of Biomedical Sciences, Academia Sinica, Taipei 11529, Taiwan. Tel: +886 2 2652 3031; Fax: +886 2 2788 7641; Email: carmay@gate.sinica.edu.tw
Correspondence may also be addressed to Tse Wen Chang at: Development Center for Biotechnology, 81 Chang Hsing Street, Taipei, Taiwan. Tel: +886 2 2735 6237; Fax: +886 2 2739 4260; Email: twchang@mail.dcb.org.tw

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors

a cDNA library of human MSPs is to subclone the cDNAs into a phage surface cDNA display system for receptor–ligand pair screening of MSPs (13,14). Despite the scientific value of a human MSP-enriched cDNA library, it is (to the best of our knowledge) not yet available in the academic community or biotechnology industry.

The goal here is to present a strategy for constructing a high-quality cDNA library that is enriched in human MSP-cDNAs. Our method is based on generating an 'optimal' set of polymerase chain reaction (PCR) *l*-nucleotide primers (l = 9, 12, 15, 18, 21) by identifying common patterns in human cDNAs encoding signal peptides (referred to as signal cDNAs). Although the primer candidates are designed to match and thus amplify DNA sequences encoding MSPs, they may also hybridize with the DNA sequences encoding non-MSPs or the DNA sequences corresponding to the mature region of MSPs to yield undesired side products, thus yielding non-MSP members in the library. Hence, noise analyses were performed to predict the products resulting from a given primer set (see Materials and Methods). The results of the noise analyses allowed us to minimize both the number of primers needed to amplify DNA encoding human MSPs and undesired products after the PCR (see Results).

MATERIALS AND METHODS

Existing databases of cDNAs encoding MSPs and non-MSPs

Human protein sequences were retrieved from the SWISS-PROT (Release 39) and TrEMBL (Release 15) databases (15) and searched for the presence or absence of signal peptides. This resulted in 1886 MSP and 22 439 non-MSP human protein sequences. For each MSP or non-MSP protein, its corresponding cDNA was obtained from the EMBL access number(s) in the SWISS-PROT and TrEMBL entries and sequences coding for proteins that include start and stop codons were selected. The cDNA sequences were translated to protein sequences and compared with the 1886 MSP and 22 439 non-MSP human protein sequences obtained directly from the SWISS-PROT and TrEMBL databases. The matches (corresponding to 100% amino acid sequence identity) resulted in 1715 and 16 138 non-redundant human cDNAs encoding MSPs and non-MSPs, respectively. The 1715 cDNAs encoding MSPs (referred to as MSP-cDNAs) do not constitute a 'complete set' as not all human MSPs are currently known (see Discussion).

Generating *l*-nucleotide primers

As primers used in PCR experiments usually contain 9–21 nt, a set of *l*-nucleotide (l = 9, 12, 15, 18, 21) primer candidates was obtained by searching for the minimal set of common patterns in the signal cDNAs. Each signal cDNA sequence was scanned by sliding windows of 9, 12, 15, 18 and 21 nt in turn. Each *l*-nucleotide pattern associated with signal cDNA sequence k (k = 1, . . ., 1715) was recorded in a hash function. The most common pattern in the pool of 1715 signal cDNAs was selected as a template for the PCR, and cDNAs containing this pattern were removed from the pool. This procedure was repeated for pools of decreasing number of signal cDNAs until no sequences were left in the pool. If a *l*-nucleotide pattern pertaining to signal cDNA sequence k is unique (i.e. it does not occur in other signal cDNAs), then its N-terminal *l*-nucleotides were selected as a primer. Our algorithm is similar to the one described by Doi and Imai (16,17).

RT–PCR to clone cDNAs encoding MSPs

First-strand cDNA synthesis can be performed by a mixture of human poly(A) + RNA with oligo-dT (reverse primer), MMLV reverse transcriptase, MMLV reaction buffer and dNTP. For the PCR reactions, the forward primers are the ones generated using the methods described in this work (see Results), whereas the reverse primers are oligo-dT comprising of 12–15 dT. The reaction conditions for amplifying different cDNAs need to be adjusted depending on the gene size and primer annealing temperatures. By predicting the MSP products stemming from each primer (see next section), restriction enzyme sites that are not present in the MSP-cDNAs could be designed, and added to both forward and reverse primers to facilitate the cloning of cDNAs into suitable library vectors (such as Lambda ZAP II vectors).

Analyses of primer-derived products

The noise levels of the selected primers in the PCR reactions were estimated by predicting the products that would result from each primer. The sequence of each primer was compared with the signal and non-signal peptide regions of the 1715 MSP-cDNAs as well as the 16 138 non-MSP-cDNAs. The resulting matches (corresponding to 100% sequence identity) were grouped according to the size (number of nucleotides) of the products. Generally, two cDNAs can easily be separated by gel electrophoresis if they differ by >200 nt. Hence, this number (200) was used to estimate the degree to which the desired and undesired products of each primer overlapped. As the longest cDNA encoding a MSP contains ~15 600 nt, respectively, the *l*-nucleotide products were grouped into 78 units; i.e. products with (one) $l \leq 200$, (two) $200 < l \leq 400$, (three) $400 < l \leq 600$, . . ., (78) $15\,400 < l \leq 156\,000$.

The statistical results of the products derived from primer i of length l were described by three parameters: (i) $N_{i,l}^{MSP}$, the number of matches between primer i of length l and the signal peptide region of MSP-cDNAs, (ii) $N_{i,l}^{MSPx}$, the number of matches between the primer and the non-signal peptide region of MSP-cDNAs, and (iii) $N_{i,l}^{nMSP}$, the number of matches between the primer and non-MSP-cDNAs. The noise fraction ($F_{i,l}$) is given by the ratio of the total number of undesired products to all possible products from a given primer i of length l, i.e.,

$$F_{i,l} = \frac{N_{i,l}^{MSPx} + N_{i,l}^{nMSP}}{N_{i,l}^{MSP} + N_{i,l}^{MSPx} + N_{i,l}^{nMSP}} \quad \mathbf{1}$$

The average noise fraction ($\overline{F}_l$) for N_l primers of length l was then obtained by averaging all $F_{i,l}$ corresponding to a fixed length, l, i.e.,

$$\overline{F}_l = \frac{1}{N_l} \sum_{i=1}^{N_l} F_{i,l} \quad \mathbf{2}$$

Estimating the number of primers needed to amplify cDNAs of all human MSPs

To estimate the number of *l*-nucleotide primers needed to amplify the cDNAs of all human MSPs, the relationship between the 'input' number of signal-cDNAs and the resulting

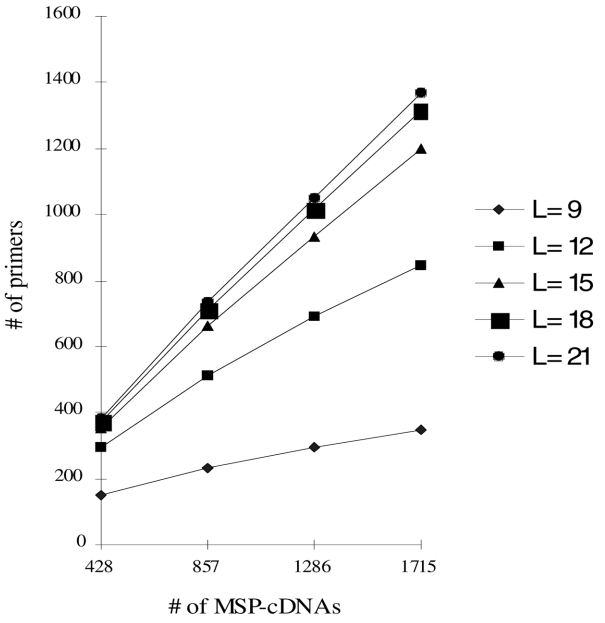


Figure 1. A plot of the average number of *l*-nucleotide primers $\bar{n}_{l,x}$ for a given *x* number of input signal cDNAs (see Materials and Methods).

number of primer candidates was obtained. First, a quarter, one-half and three-quarters of the 'full' set of 1715 MSP-cDNAs were randomly selected to yield subsets of 428, 857 and 1286 input sequences, respectively (referred to as subsets $S_{1/4}$, $S_{1/2}$ and $S_{3/4}$). Ten subsets each of $S_{1/4}$, $S_{1/2}$ and $S_{3/4}$ MSP-cDNAs were randomly generated to yield a total of 30 input data sets. For each subset of MSP-cDNAs, *l*-nucleotide primer candidates were obtained by scanning each MSP-cDNA sequence using sliding windows of 9–21 nt as described above. The total numbers of *l*-nucleotide primers corresponding to the 10 subsets S_x were then averaged to yield $\bar{n}_{l,x}$, which were fitted to various (linear, quadratic, cubic and power) functions of the subset size *x*, i.e., $\bar{n}_{l,x} = f_l(x)$ (Fig. 1). The resulting function allowed us to obtain an upper bound on the number of *l*-nucleotide primers needed to amplify the cDNAs of all human MSPs.

RESULTS

The maximal number of *l*-nucleotide primers needed to amplify all human MSP-cDNAs

Figure 1 plots the average number of *l*-nucleotide primers $\bar{n}_{l,x}$ for a given *x* number of input signal cDNAs (see Materials and Methods). The relative standard deviation (standard deviation/average number of primers) of each $\bar{n}_{l,x}$ data point in Figure 1 is ≤3.7% (Table 1). This means that a similar number of primers can amplify sets of size *x* containing different MSP-cDNAs. A quadratic function was found to best fit $\bar{n}_{l,x}$ as a function of *x* for primers of a given length *l*. This functional form allowed us to obtain an upper bound on the number of *l*-nucleotide primers that is needed to amplify the cDNAs of all human MSPs. If the latter is 4000 (see Introduction), then the maximal number of primers of length 9, 12, 15, 18 and 21 is estimated to be 366, 1192, 2207, 2578 and 2684, respectively.

Table 1. Relative standard deviation (standard deviation/average number of primers) of each $\bar{n}_{l,x}$ in Figure 1

No. of MSP-cDNAs[a]	428	857	1286
l = 9	3.7	1.7	1.8
l =12	2.8	2.0	0.9
l =15	1.9	1.5	0.8
l =18	1.9	1.1	0.5
l = 21	1.5	1.0	0.6

[a]1/4, 1/2 and 3/4 of the 'full' set of 1715 MSP-cDNAs were randomly selected to yield subsets of 428, 857 and 1286 input sequences respectively (see Materials and Methods).

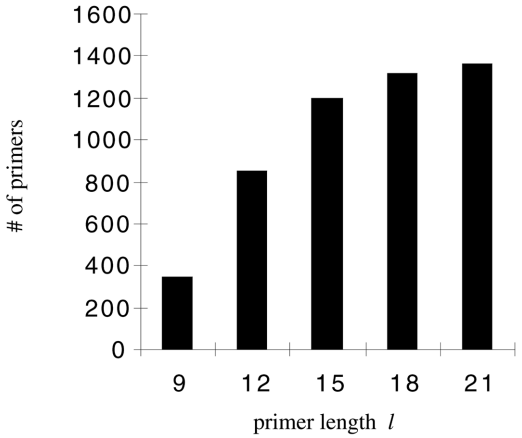


Figure 2. The 'minimal' number of *l*-nucleotide primers needed to amplify signal cDNAs of human MSPs as a function of the primer length *l*.

Dependence of the primer set size and noise level on primer length

Figure 2 shows that the 'minimal' number of *l*-nucleotide primers needed to amplify human MSP-cDNAs increases with increasing primer length. It is 350, 848, 1198, 1317 and 1366 for primers containing 9, 12, 15, 18 and 21 nt, respectively. However, the number of human non-MSP-cDNAs derived from the *l*-nucleotide primers decreased dramatically with increasing primer length. This is illustrated in Figure 3, which shows that the 9 nt primers yield >97% cDNAs encoding mature regions of MSPs and non-MSPs, but the longest primers generate <2% undesired products. The opposite effects of the primer set size and average noise fraction on the primer length imply that it is not possible to obtain a low noise level using a small set of short primers to amplify the cDNAs of human MSPs.

Table 2 shows the noise fraction distribution for sets of primers varying in lengths from 9 to 21. All of the 9-nt primers generated a significant fraction of undesired products. A high percentage (86%) of the 12-nt primers also generated >10% undesired products. The noise level was dramatically improved with the longer primers as only 16, 5 and 4% of the 15-, 18- and 21-nt primers, respectively, produced >10% undesired products. The lower noise levels found for the longer primers should facilitate product separation. This is verified in Figure 4, which shows the average product distribution (i.e. the

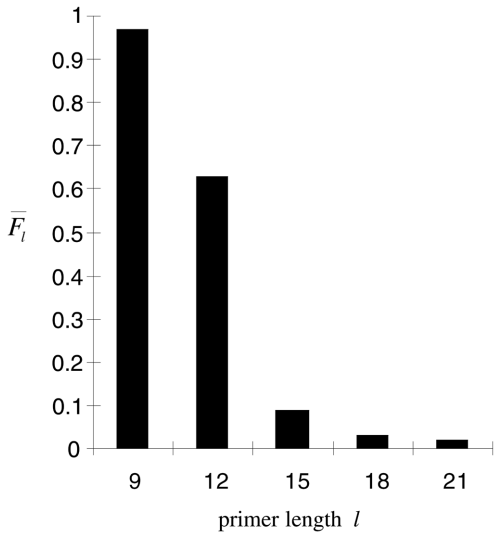


Figure 3. A plot of the average noise fraction, $\overline{F}_l$ (equation **2**), as a function of the primer length *l*.

Table 2. Distribution of noise fraction $F_{i,l}$ for primers of length *l*[a]

Primer length	9	12	15	18	21
No. of primers	350	848	1198	1317	1366
No. of primers in N-region	55	478	1024	1227	1301
No. of primers in H + C region	295	370	174	90	65
No. of degenerate primers	40	21	1	0	0
$0.0 \le F_{i,l} < 0.1$	0.0	14.2	83.7	94.9	95.9
$0.1 \le F_{i,l} < 0.2$	0.0	0.2	0.2	0.1	0.1
$0.2 \le F_{i,l} < 0.3$	0.0	0.7	1.2	0.4	0.2
$0.3 \le F_{i,l} < 0.4$	0.0	3.1	2.1	1.0	0.7
$0.4 \le F_{i,l} < 0.5$	0.3	1.3	1.0	0.3	0.4
$0.5 \le F_{i,l} < 0.6$	0.0	14.3	6.5	2.2	2.0
$0.6 \le F_{i,l} < 0.7$	0.0	10.8	2.6	0.6	0.4
$0.7 \le F_{i,l} < 0.8$	0.0	16.7	1.8	0.4	0.3
$0.8 \le F_{i,l} < 0.9$	2.9	26.3	0.8	0.2	0.0
$0.9 \le F_{i,l} < 1.0$	96.9	12.4	0.0	0.0	0.0

[a]The numbers in columns two to six are the percentage numbers of primers with $F_{i,l}$ in the range specified in column one.

number of products in a given size range divided by the number of *l*-nucleotide primers) as a function of the product size (see Materials and Methods). As Figure 4A shows very few MSP-cDNAs longer than 4000 nt, only products containing ≤4000 nt are shown in Figure 4B–F for the sake of clarity. Figure 4B and C show that it is difficult to separate the MSP-cDNAs derived from 9- and 12-nt primers, whereas Figure 4D–F shows that it is possible to isolate the MSP-cDNAs stemming from primers comprising of ≥15 nt. The results in Table 2 combined with those in Figures 2–4, indicate that the 21-nt primers are expected to make a higher quality library (with less undesired products), but they would be more expensive as the cost of a primer is proportional to its length.

Primer composition and features

One way in which the number of 21-nt primers could be reduced is to identify degenerate primers such as NTGCTGCTG, where N denotes any of the four bases. Table 2 shows that while degenerate sequences are found in primers <18 nt long, they are absent from the longer primers. Table 2 also shows that the shortest primers are derived mostly from the H + C region, whereas the longer ones pertain mostly to the N-region (see Introduction). The latter is a consequence of the primer generation procedure employed (see Materials and Methods). Consider a pool of three sequences, A, B, C, composed of P1 + P2 + P3, P4 + P5 + P2 and P6 + P7 + P3, respectively (where Pn denotes a given pattern/primer). Among the three sequences P2 is the most common pattern, hence it will be selected as a primer candidate and sequences A and B will be removed from the pool. As the patterns in the remaining C sequence are unique, the N-terminal primer (P6) will be selected to amplify C, even though P3 is shared by both C and A. Table 3 lists the top 10 most common *l*-nucleotide primers with noise fraction ($F_{i,l}$) values <0.1. The corresponding amino acid sequence motifs/patterns comprise mostly of hydrophobic residues in the H region (see Introduction). They appear to be characteristic of MSPs and can help to identify novel MSPs that contain these signal sequences.

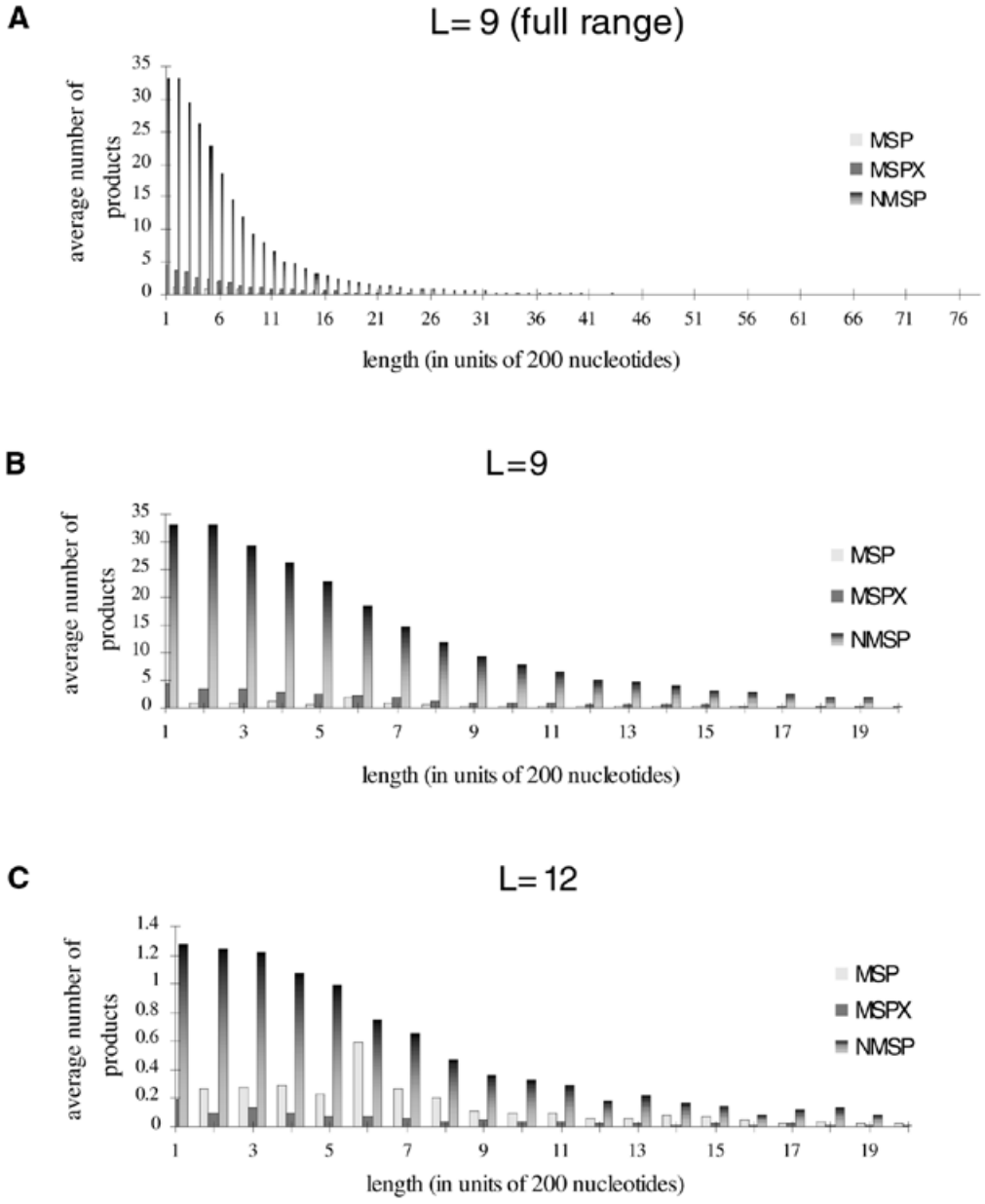
Selecting an optimal set of mixed-length primers

To achieve a balance between quality (low noise level) and cost (length of primer), the number of 21-nt primers (1366) was reduced in the following manner. First, the noise level limit was set equal to *x*. All shorter ($l < 21$) primers whose $F_{i,l} \le x$ were mixed with the 21-nt primers to constitute a candidate set of patterns (see Materials and Methods). The most frequently occurring *l*-nucleotide pattern/primer in the pool of 1715 signal cDNAs was chosen as a template for the PCR and cDNAs containing this *l*-nucleotide primer were removed from the pool. This procedure was repeated for pools of decreasing number of signal cDNAs till no sequences were left in the pool.

Table 4 shows that when the noise level limit was set equal to 0.1, the number of 21-nt primers (1366) could be reduced without sacrificing the quality of the library. As for the 21-nt primers, only ~4% of the 1302 mixed-length primers generated >10% undesired products. When the noise level threshold was raised, the number of mixed-length primers decreased, but the percentage number of primers generating >10% undesired products increased. For example, with a noise level limit of 0.5, the number of mixed-length primers was 1203, but 22% of these primers generated >10% undesired products. The results in Table 4 suggest using a noise threshold of 0.1 to derive an 'optimal' set of 1302 mixed-length primers for amplifying human MSP-cDNAs.

Modifying primers for RT–PCR

Some of the primer candidates should be modified before application. The primers can be separated into groups N, H and C, corresponding to the N-region, H-region and C-region of the signal peptides, respectively (see Introduction). Most of the *l*-nucleotide primers that can amplify more than one MSP-cDNA belong to group H (see above and Table 3). Primers belonging to the H or C group will generate products with non-functional signal peptides, hence they need to be



modified. Previous work showed that a 'minimal' signal peptide in eukaryotes could be constructed out of a single positively charged N-terminal residue following the start codon, a seven-residue H-region and a five-residue C-region (4). Thus, the N and N + H regions of the 'minimal' signal cDNA sequence should be added at the 5′ end to primers belonging to the H and C groups, respectively. On the other hand, group N primers can be cloned directly into a suitable vector except that ATG should be added at the 5′ end to primers without an initiator methionine (5′ ATG).

DISCUSSION

The results above suggest using a set of mixed-length primers (obtained using a noise threshold of 0.1) to construct a high-quality cDNA library enriched in human MSPs (Figs 2 and 3). These primers generated ≤4% undesired products (Table 4), and the desired MSP-cDNA products could be easily separated (Fig. 4). Furthermore, the number of PCR reactions that need to be carried out could be reduced using the multiplex PCR, in which multiple sets of primers are used to amplify multiple genes in a single reaction. For example, 18 non-complementary forward and reverse primers had been used to amplify nine different exons in a single PCR reaction (18). Thus, the 'forward' primers (see Materials and Methods) may be divided into groups, each containing nine or fewer non-complementary primers. This could potentially reduce the number of required PCR reactions.

As not all genes encoding human MSPs are known at present, the methodology outlined in this work could be used to identify novel MSPs. This is based on the finding that many primers with low noise fraction ($F_{i,l} \leq 0.1$) occur in more than

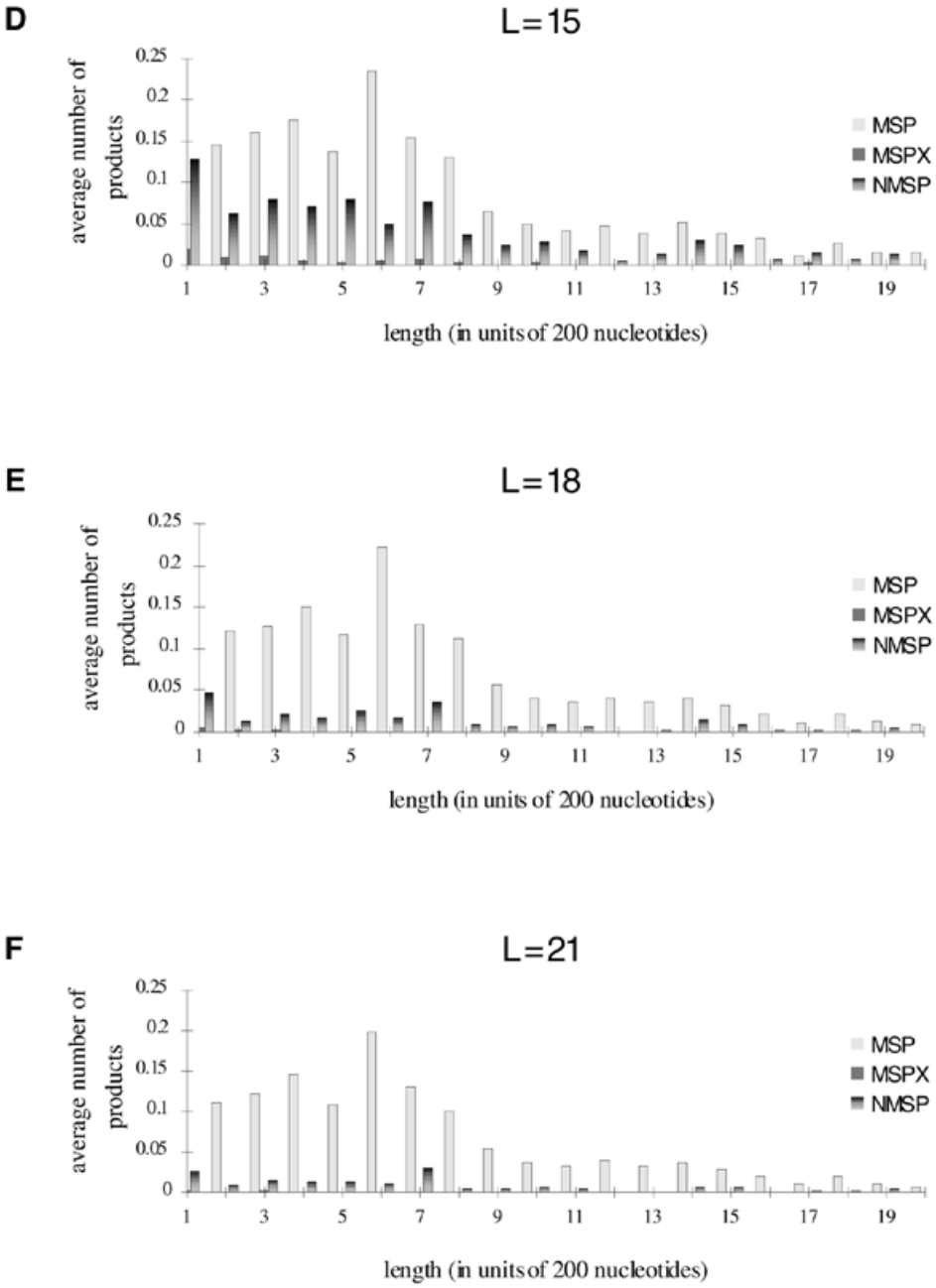


Figure 4. (Previous page and above) The average product distribution (i.e. the number of products in a given size range divided by the number of *l*-nucleotide primers) as a function of the product size (see Materials and Methods).

one MSP. Hence if a primer in our 'optimal' primer set shares part of the signal cDNA sequence of a novel MSP, then it will generate that MSP. To verify that the resulting product is indeed novel, its sequence could be aligned with the known MSP-cDNAs. Furthermore, to verify that it is a MSP, its protein sequence could be examined for the presence of a signal peptide and its cleavage site using signal peptide prediction programs (19,20; C.Y.Wu and C.Lim, manuscript in preparation). These programs can also be used to screen the putative protein products corresponding to cDNAs amplified by a given primer to avoid mistaking desired MSP-cDNAs for noise.

The several methods for preparing cDNA libraries of MSPs described in the introduction all have their basis for encompassing proportions of the MSPs in the constructed libraries. The previous methods took advantage of certain cellular or biochemical properties of the MSP-mRNAs. For example, the method employing membrane-bound polysomes as the source of mRNAs for MSPs is based on the fact that the MSP-mRNAs are translated on ribosomes attached to the surface of endoplasmic reticulum. Hence, this method may misclassify some mRNAs encoding cytosolic proteins as MSPs if they associate with cellular membranes (10). Other methods such as the selection process using defective invertase gene utilize a

Table 3. The top 10 most common *l*-nucleotide primers with $F_{i,l} < 0.1$

Nucleotide code	Amino acid	Region	$F_{i,l}$
Gcg ccc cga acc	APRT	H	0.03
Ctg acc gag acc tgg	LTETW	H + C	0.01
Ggg gcc ctg gcc ctg	GALAL	H	0.07
Gcc ctg acc gag acc tgg	ALTETW	H + C	0.01
Ggg gcc ctg gcc ctg acc	GALALT	H + C	0.04
Atg cag ccg agg tgg gcc	MQPRWA	N	0.00
Gcc ctg acc gag acc tgg gcc	ALTETWA	H + C	0.02
Gcg ccc cga acc ctc ctc ctg	APRTLLL	H	0.00
Ctg gcc ctg acc cag acc tgg	LALTQTW	H + C	0.05
Atg cag ccg agg tgg gcc caa	MQPRWAQ	N	0.00
Atg gcc ctg tcc ttt tct tta	MALSFSL	N	0.00

Table 4. Distribution of noise fraction $F_{i,l}$ for mixed-length primers[a]

Noise threshold[b]	0.1	0.2	0.3	0.4	0.5
No. of primers	1302	1295	1286	1240	1203
$0.0 \le F_{i,l} < 0.1$	95.8	95.2	94.2	89.8	78.0
$0.1 \le F_{i,l} < 0.2$	0.1	0.4	0.3	0.3	0.3
$0.2 \le F_{i,l} < 0.3$	0.2	0.5	1.6	1.4	1.2
$0.3 \le F_{i,l} < 0.4$	0.7	0.7	0.5	4.4	4.3
$0.4 \le F_{i,l} < 0.5$	0.4	0.4	0.4	1.2	1.8
$0.5 \le F_{i,l} < 0.6$	2.1	2.1	2.0	2.1	13.5
$0.6 \le F_{i,l} < 0.7$	0.5	0.5	0.5	0.5	0.5
$0.7 \le F_{i,l} < 0.8$	0.3	0.3	0.3	0.3	0.3
$0.8 \le F_{i,l} < 0.9$	0	0	0	0	0
$0.9 \le F_{i,l} < 1.0$	0	0	0	0	0

[a]The numbers in columns two to six are the percentage numbers of primers with $F_{i,l}$ in the range specified in column one.
[b]The noise level limit used to replace the 21-nt primers with shorter primers (see Results).

selection molecular process to channel the mRNA carrying the secretion signal sequences to be secreted or expressed. The invertase genetic selection in yeast may not detect certain human MSPs if the human signal peptide cannot direct an invertase enzyme to the yeast secretory pathway, which is similar but not identical to the human secretory pathway (1,2).

The present approach, also aimed at obtaining a cDNA library encompassing a large proportion of MSPs, is different from the earlier approaches in that it is based on the development in genomic research and the availability of DNA sequences of a portion of MSPs. The availability of these cDNAs serves two purposes in our design of the cDNA library. First, we have a means to ensure the completeness in our design of primers, which is based on the information of every individual member in the database. Secondly, the sequence information for the known MSPs enables us to design primers that can statistically cover some unidentified MSPs in the genome. This, therefore, helps us to identify MSPs and study their functions. Establishing an MSP-enriched cDNA library would also have a positive impact on protein sub-cellular location prediction, as discussed in a recent paper (21). Furthermore, when public databases for all human cDNAs become available, a complete set of all human MSP-cDNAs can be obtained, which will yield an 'optimal' set of mixed-length primers employing the strategy outlined here. This set of primers, in principle, would generate a library enriched in the cDNAs of all human MSPs.

As different sets of MSPs are expressed in different proportions by different tissues, the mRNA isolated from a tissue or a cell line contain only a portion of the mRNA of MSPs that are formed in the entire human body. The same tissues at normal or disease states may also express certain MSPs differently. The same cells may express certain MSPs differently at different stages in the cell cycle or at different stages of differentiation. The primers designed by the present study should be applicable to prepare cDNA libraries of various tissues. For some applications, the mRNA from different tissues should be amplified separately. For other applications, mRNA from certain tissues may be pooled to prepare a pooled cDNA library.

ACKNOWLEDGEMENTS

We thank Drs Pei-ing Hwang, Konan Peck, Jim Sheu and Adam Yao for helpful suggestions on the manuscript. This work is supported by the Institute of Biomedical Sciences at Academia Sinica and the National Science Council, Taiwan.

REFERENCES

1. Klein,R.D., Gu,Q., Goddard,A. and Rosenthal,A. (1996) Selection for genes encoding secreted proteins and receptors. *Proc. Natl Acad. Sci. USA*, **93**, 7108–7113.
2. Jacobs,K.A., Collins-Racie,L.A., Colbert,M., Duckett,M., Golden-Fleet,M., Kelleher,K., Kriz,R., La Vallie,E.R., Merberg,D., Spaulding,V. *et al.* (1997) A genetic selection for isolating cDNAs encoding secreted proteins. *Gene*, **198**, 289–296.
3. Goo,J.H., Park,A.R., Park,W.J. and Park,O.K. (1999) Selection of *Arabidopsis* genes encoding secreted and plasma membrane proteins. *Plant Mol. Biol.*, **41**, 415–423.
4. Von Heijne,G. (1985) Signal sequences: The limit of variation. *J. Mol. Biol.*, **184**, 99–105.
5. Claros,M.G., Brunak,S. and Von Heijne,G. (1997) Prediction of N-terminal protein sorting signals. *Curr. Opin. Struct. Biol.*, **7**, 394–398.
6. Tashiro,K., Tada,H., Heilker,R., Shirozu,M., Nakano,T. and Honjo,T. (1993) Signal sequence trap: A cloning strategy for secreted proteins and type I membrane proteins. *Science*, **261**, 600–603.
7. Chen,H. and Leder,P. (1999) A new signal sequence trap using alkaline phosphatase as a reporter. *Nucleic Acids Res.*, **27**, 1219–1222.
8. Peterfy,M., Gyuris,T. and Takacs,L. (2000) Signal-exon trap: a novel method for the identification of signal sequences from genomic DNA. *Nucleic Acids Res.*, **28**, e26.
9. Kopczynski,C.C., Noordermeer,J.N., Serano,T.L., Chen,W.Y., Pendleton,J.D., Lewis,S., Goodman,C.S. and Rubin,G.M. (1998) A high throughput screen to identify secreted and transmembrane proteins involved in *Drosophila* embryogenesis. *Proc. Natl Acad. Sci. USA*, **95**, 9973–9978.
10. Diehn,M., Eisen,M.B., Botstein,D. and Brown,P.O. (2000) Large-scale identification of secreted and membrane-associated gene products using DNA microarrays. *Nature Genet.*, **25**, 58–62.
11. Ladunga,I. (2000) Large-scale predictions of secretory proteins from mammalian genomic and EST sequences. *Curr. Opin. Biotechnol.*, **11**, 13–18.

12. Venter,J.C., Adams,M.D., Myers,E.W., Li,P.W., Mural,R.J., Sutton,G.G., Smith,H.O., Yandell,M., Evans,C.A., Holt,R.A. *et al.* (2001) The Sequence of the Human Genome. *Science*, **291**, 1304–1351.
13. Crameri,R., Hemmann,S. and Blaser,K. (1996) PJuFo:a phagemid for display of cDNA libraries on phage surface suitable for selective isolation of clones expressing allergens. *Adv. Exp. Med. Biol.*, **409**, 103–110.
14. Crameri,R. and Walter,G. (1999) Selective enrichment and high-throughput screening of phage surface-displayed cDNA libraries from complex allergenic systems. *Comb. Chem. High Throughput Screen*, **2**, 63–72.
15. Bairoch,A. and Apweiler,R. (1999) The SWISS-PROT protein data bank and its supplement TrEMBL in 1999. *Nucleic Acids Res.*, **27**, 49–54.
16. Doi,K. and Imai,H. (1997) Greedy algorithms for finding a small set of primers satisfying cover and length resolution conditions in PCR experiments. *Genome Inform. Ser. Workshop Genome Inform.*, **8**, 43–52.
17. Doi,K. and Imai,H. (1999) A greedy algorithm for minimizing the number of primers in multiple PCR experiments. *Genome Inform. Ser. Workshop Genome Inform.*, **10**, 73–82.
18. Mullis,K.B., Ferre,F. and Gibbs,R.A. (1994) Optimization of Multiplex PCRs. *The Polymerase Chain Reaction*. Birkhauser, Boston, pp. 38–46.
19. Nielsen,H., Brunak,S. and Von Heijne,G. (1999) Machine learning approaches for the prediction of signal peptides and other protein sorting signals. *Protein Eng.*, **12**, 3–9.
20. Chou,K.C. (2001) Using subsite coupling to predict signal peptides. *Protein Eng.*, **14**, 75–79.
21. Chou,K.C. (2001) Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins*, **43**, 246–255.