GeneUp: A Program to Select Short PCR Primer Pairs that Occur in Multiple Members of Sequence Lists

BioTechniques 25:112-123 (July 1998)

G. Pesole, S. Liuni[1], G. Grillo[2], P. Belichard[3], T. Trenkle[3], J. Welsh[3] and M. McClelland[3]
Universitá della Basilicata, Potenza, [1]Centro di Studio sui Mitocondri e Metabolismo Energetico, Bari, [2]Universitá di Bari, Bari, Italy and [3]Sidney Kimmel Cancer Center, San Diego, CA, USA

ABSTRACT

A computer program is presented that selects a small set of short primer pairs for PCR to sample all the sequences in a user-specified list of mRNAs. Such primer pairs could be used to increase the probability of sampling mRNAs of particular interest in differential display and to generate simplified hybridization probes for DNA chips or arrays. The program uses simulated PCR to find pairs of primers that sample more than one sequence in the list. A small set of such primer pairs is selected that give maximal coverage of the sequences in the list. Primer pairs are excluded that: (i) *generate simulated PCR products of the same size from a number of sequences in the list,* (ii) *can easily form primer dimers,* (iii) *are outside a specified range of G+C content or* (iv) *occur in another list of undesirable sequences, such as rRNAs and Alu repeats. Five lists consisting of from 48–285 cDNA sequences were used to test the program. A small number of pairs of primers, 8–10 bases in length, were selected that fit the above criteria and that generate one or more simulated PCR products in all or most of the cDNAs in each list.*

INTRODUCTION

The polymerase chain reaction (PCR) uses specific DNA primers to amplify specific sequences from a complex source of nucleic acids (13). If arbitrarily selected primers are used instead, then a reproducible fingerprint of products can be generated under the appropriate conditions (8,20,21). Arbitrary primers initiate PCR from sites on the template with varying efficiencies depending on the quality of the overall match and with particular regard to the match at the 3′ end of the primer. Relatively efficient priming events within a few thousand base pairs and facing each other on opposite strands, lead to PCR-amplifiable products. Products that are the most efficiently primed and most efficiently amplified compete most effectively in the subsequent PCR amplification and are visualized as a fingerprint after gel electrophoresis.

This method was first applied to detect polymorphisms among related genomes because differences in primer-binding sites result in differences in the resulting PCR products (21,26). Later, the method was applied to studying differential expression of arbitrarily sampled RNAs (8,20). Differences in a cDNA PCR product derived from two isogenic RNA samples reflected differences in the abundance of the mRNA between the populations.

One of the features of the method that could be changed is the arbitrary nature of the sampling. The first efforts to direct PCR fingerprints to particular sequences were applied to genomic DNAs. For example, the rate of detection of polymorphisms in higher eukaryote nuclear DNA could be increased by using primers that target the more polymorphic simple repeats or hyper-mutable methylation sites (25). Primers can also be directed towards sequences that are more frequent in some bacterial genomes (12,22,24,27).

PCR with pairs of degenerate primers derived from back translation of conserved amino acid motifs have been widely used for finding new members of gene families (e.g., Reference 3). Similarly, PCR fingerprinting with primers derived from conserved motifs sometimes enriches for genes of interest (4,18,28). Another approach to selecting primers for targeted fingerprinting is to determine which oligonucleotide sequences are common in the list of sequences of interest and determine which pairs of primers will give a significant number of PCR products from these sequences. With the addition of arbitrary 5′ tails, these primers can then be used at relatively high stringency to sample sequences from the family of interest; although, other mRNAs could also be sampled by the same primers (9).

We present a program to effectively select perfectly matched primer pairs for all or most of the sequences in a list of interest. Primers that match hyper-abundant RNAs are excluded.

RATIONALE

Previously, at least two programs have been written that select primer pairs that generate a simulated PCR product

Table 1. GenBank Accession Nos. of DNAREP Sequences Used in This Study

(DNAREP)

1	D64108	2	X83441	3	L24444	4	X91992	5	D13370
6	Z30094	7	L27425	8	X84740	9	M28650	10	X52221
11	M31899	12	L04791	13	D21235	14	D21090	15	M74524
16	M74525	17	M36067	18	X69821	19	M29971	20	L47579
21	X81030	22	D38500	23	D38501	24	D38502	25	D38503
26	D13804	27	L33262	28	X97795	29	M29474	30	M94633
31	L07872	32	X78627	33	X78262	34	U61981	35	U63139
36	U63329	37	U64315	38	D14533	39	Z11495	40	D21089
41	X69978	42	X71342	43	L34079	44	K03199	45	L09561
46	L37374	47	M87499	48	U09559	49	U12134	50	U13695
51	U13696	52	U18300	53	U27346	54	U28946	55	U32986
56	U37359	57	U40622	58	U40671	59	U47077	60	U72936
61	U75967	62	X15653	63	X83753	64	Y10658	65	Z48796

(DNAREP): 65 human DNA repair-associated mRNAs. Before CLEANUP application, the collections contained 169 sequences.

from more than one sequence in a list of sequences of interest (9,14). In one of these programs (9), primers are chosen that have a frequency above a set threshold in the sense and antisense strands among the list of sequences of interest. Then these primers are picked into random groups of thirty, which are used to generate simulated PCR products. After thousands of iterations of this sampling strategy, those primers that are most often found in groups of primers that performed “well” are chosen for a final list of primers. Then, a particular primer pair was chosen by the authors in Reference 9 that was shown to successfully target some mRNAs of interest under relatively high stringency conditions.

The ability to use these primers for biological experiments has been demonstrated. Thus, we set out to determine if we could not only develop a method to select primer pairs that would also sample multiple sequences in a list, but also allow maximum coverage of sequences in the list while excluding primer pairs that could be a problem because they might dominate the resulting mixture of PCR products. To achieve this goal, we decided to select primers (and their complements) that occurred frequently in either strand of the sequences of interest and to use these highly ranked primers in every pairwise combination to generate a matrix of simulated PCR products. Then, by ranking primer pairs by the number of PCR products they generated and sequences by the number of primer pairs that generated a simulated PCR product, we could systematically pick primer pairs by working across and down the matrix to get maximum coverage, as will be described later. This “greedy” algorithm uses a strategy similar to that used by Pearson et al. (14).

To test the program, we developed four sets of human cDNAs and fragments of cDNAs. Two of the lists comprised cDNAs for phylogenetically unrelated proteins that participated in particular kinds of functions in the cell. One of these lists comprised cDNAs for genes involved in DNA repair and replication (DNAREP), and the other list comprised cDNAs that are known or suspected to be associated with apoptosis (APO). The other two lists were of phylogenetically related cDNAs. One list was human nuclear receptor (HNR) cDNAs (10). The final list was human G-protein-coupled receptors (GPCR), a family that had previously been used to select primer pairs using a program of a different design (9), so we could compare the performance of our program.

PROGRAM DESIGN AND RESULTS

The program was written in the C language on a UNIX™ operating system. This program, GeneUP, and a manual are available to noncommercial users (contact **mmcclelland@skcc.org**). To remove duplicated sequences, we applied the CLEANUP program to each collection (5) (available at **ftp://area.ba.cnr.it/pub/software/cleanup**), setting the parameters to remove the shorter of any two sequences overlapping by 95% with a similarity higher than 95%. Also, sequences with a length of <800 bases were removed from each collection. While it is difficult to select primer pairs that sample multiple sequences of <800 bases, users have the option to attempt to sample all sequences in their list, regardless of length. The final set of cDNAs for the DNAREP test list is presented in Table 1. For the other lists, see **ftp://ftp. skcc.org/mcclelland/geneupm.cl**. Before and after CLEANUP, the length of each list was as follows: (DNAREP), 169 vs. 65 human DNA repair-associated mRNAs; (APO), 181 vs. 59 human apoptosis-associated mRNAs; (HNR), 62 vs. 48 human nuclear receptor mRNAs; and (GPCR), 206 vs. 113 human G-protein-coupled receptor mRNAs.

We then set out to devise a strategy that would yield the most useful primer pairs that would sample these genes in a PCR strategy. The WORDUP program (15) was adapted to

Table 2. GenBank Accession Nos. of Hyper-Abundant Sequences Used to Exclude Primers in This Study

Alu elements:					
U14567	U14568	U14569	U14570	U14571	U14572
U14573	U14574				
LINE Elements and 3′ mRNAs fragments carrying parts of LINE elements:					
AA017128	AA018943	AA022026	AA055654	AA057222	AA074788
AA076364	AA081957	AA081993	AA082639	AA084139	AA084303
AA085706	AA085707	AA088273	AA088381	AA095194	AA100276
AA102000	AA112088	AA112316	AA112323	AA113978	AA121767
AA121839	AA121840	AA121875	AA121876	AA121916	AA126794
AA126847	AA128621	AA128858	AA129985	AA130476	AA130536
AA131481	AA136629	AA136637	AA136721	AA136934	AA136977
AA148366	C17235	D58460	H03599	H13052	H20876
H82488	H85238	H92806	M78222	M85371	N20521
N22643	N23244	N23646	N23655	N23657	N23864
N24958	N25053	N29555	N33076	N41014	N70045
N76123	N76330	N79938	R14820	T02866	T02882
T03057	T03259	T06602	T06958	T07197	T16214
T48647	T56669	T57073	T57474	T57745	T59577
T60735	T63720	T90251	W03161	W03511	W19702
W26931	W26997	W27003	W37681	W58442	W85828
W90097	W90195	W93703			
K03432|HUMRGEA Human 18S rRNA gene					
M11167|HUMRGM Human 28S ribosomal RNA gene					
J01866|HUMRRB Human 5.8S ribosomal RNA					
D38112|HUMMTA Human mitochondrial DNA, complete sequence					

generate a list of the most common oligomers of various lengths from each of these four lists, counting the occurrence of the oligonucleotides in both strands.

The first challenge was to determine the appropriate length of the primers. In a random distribution of all four nucleotides, a typical 10-mer sequence occurs about once every 1 000 000 bp (4^{10}). However, sequences in biological samples rarely approach this random model. In long lists of cDNAs, the most frequent 10-mers may occur often enough to be useful as primers. However, in the shorter lists of cDNAs used in the examples we present, even the most common 10-mers occur only a few times in the four lists, and 11-mers were generally confined to regions of conserved amino acid motifs (data not shown). Thus, 9-mers or shorter might be required for these shorter lists.

In a random DNA sequence, a typical 9-mer has a frequency of about one in 131 000 bp (most of which do not occur even once in all 285 mRNAs in all four lists). However, the most prevalent 9-mer occurs 22 times in the DNAREP list, 17 times in the APO list, 20 times in the HNR list and 40 times in the GPCR list. Similarly, a typical 8-mer has a frequency of one in 32 000 bp (about three times in a list of 50 mRNAs), and the most prevalent 8-mer occurs 36 times in the DNAREP list, 30 times in the APO list, 34 times in the HNR list and 68 times in the GPCR list.

Table 3: Effect of Removing Primers that Occur in Hyper-Abundant RNAs

	Primers Remaining	
List	8-mers	9-mers
DNAREP	15	38
APO	15	30
HNR	26	50
GPCR	28	45

The-top ranked 100 primers were examined, and those that occurred in the hyper-abundant RNAs were removed.

Gresshoff and colleagues have shown that primers as short as five bases could be used for PCR (2). However, short primers of 5 or 6 bases, while they would occur frequently in the list of interest, would also occur frequently in other RNAs and therefore might not generate any significant selectivity for the mRNAs in our list of interest.

Although primers of 8–12 bases in length are probably the most useful for the applications described here, the program has the option to search for oligomers from 6-mers to any length in a given list of sequences (with limitations to the number and length of sequences to be analyzed dependent on the memory available on the computer).

The program can exclude primers that occur in a user-specified list of sequences. Specifically, we wished to exclude primers that occurred in hyper-abundant RNAs, such as ribosomal RNAs, mitochondrial RNAs and dispersed repeats, on the theory that cDNA from such sources could adversely affect the results by leading to a few dominant PCR products. In human RNAs, the most abundant dispersed repeats are Alu and LINE (1,7,16). Even after poly(A) selection of mRNA or oligo(dT) priming, Alu and LINE repeats are undiminished, and rRNAs and mtRNAs generally still constitute a substantial but variable minority of the resulting population. A substantial fraction of 9-mers and the majority of 8-mers can be expected to occur in this list of hyper-abundant RNAs. Exclusion of perfect matches in these RNAs should reduce cDNA synthesis and subsequent arbitrary priming in these hyper-abundant RNAs and thus improve targeting to the less-abundant RNAs of interest during PCR amplification.

Note that primers that have only a single mismatch in a hyper-abundant RNA may still prime quite efficiently on that undesirable target. As a step towards removing such primers, we have added an option for users to remove all 10-mers that have a perfect match with a hyper-abundant 9-mer at the 3′ end or, in general, all *n*-mers that have a perfect match with a hyper-abundant (*n*-1)-mer at the 3′ end.

In the examples we present here, primers are subtracted that occur in a list of the rRNAs, mtDNA and eight representative Alu elements, compiled in Table 2. Every occurrence of a potential primer match in either strand of these sequences is excluded. In addition, a list of 99 mRNAs that carry fragments of LINE elements was compiled. LINE element sequences are most typically found in a 5′ truncated form in the 3′ end of mRNAs (6). Any oligonucleotide that occurs three or more times in either strand in this list is excluded from the oligomers generated from the list of sequences of interest. This step removed oligonucleotides that were common in LINE elements as well as some that were common in the non-coding 3′ ends of these 99 mRNAs. Only oligonucleotides over 7 bases in length remain after the subtraction step because the vast majority of the 16 384 possible 7-mers occur in rRNAs, mtDNA, Alu or in the 3′ end of LINEs. Thus, we confined further analyses to lists of 8-mer and 9-mer oligonucleotides. Table 3 shows the effect of subtract-

(A)

Primer Pair		Primer Sequences (5′ to 3′)
D8′-C8	:	GGAAGGAG-CTCCTGCA
T9-B17	:	CTGGCTGA-TGAGGAAG
E7-M14	:	CCTCCTGG-GGGGCAGC
U18′-T10′	:	AGGAGGAA-CTCCTTTC
J12-W6	:	GCCAGTGG-CCCAGCCT
U6′-J0′	:	GCTCTGGG-CAGGCTGT
T9′-N15	:	TCAGCCAG-TCAGGAAG
D8′-G11′	:	GGAAGGAG-CCTGGCTC
I10′-S16	:	AGCTGAAG-TCTGGGCT
K5-S7′	:	CATCCAGA-TCTGAAGG
D5′-K1	:	AGAACCTG-AGAGCTTC
B11-S15	:	GAGAAGCA-TCCAGCAG
B5-B10	:	CAGGTGGA-CTGTCACC
Q2-M10	:	AGTTCCTG-CTTCTGGA
Z0′-H7	:	GCTGAAGT-CCTCTGCT
Y4′-Z1	:	CAGCCCTG-AGCCCCAG
T7′-P10	:	GCTGAAGG-CTTGCTGA
I12′-X10′	:	GGCCTGGC-GCCACTTC
S4-O15	:	CAGGAGAG-TCAGGAGG
S11-F5	:	GCAGAAGT-CAGTGGTG
Z5-C4	:	CCAGAAGG-CAGCAGAG
E17-D14	:	TGCAGATG-GGCTCTCT

(B)

Primer Pair	Gene No.	Product Length
D8′-C8	17	121
	27	522
	34	154
	36	682
	40	138
	43	742
	49	522
	64	355
T9-B17	5	526
	8	122
	29	541
	45	247
	47	496
	58	122
E7-M14	2	422
	10	435
	40	163
	44	266
	61	547
	65	191
U18′-T10′	1	904
	17	305
	43	481
	53	177
	59	666
	61	490
J12-W6	3	382
	8	393
	21	237
	40	955
	58	393
	59	231
U6′-J0′	22	433
	23	431
	24	540
	36	359
	51	561
	55	941
T9’-N15	8	51
	26	251
	58	51
	60	276
	63	276
D8′-G11′	6	816
	7	96
	17	228
	27	798
	49	798
I10′-S16	12	603
	27	720
	37	234
	49	720
	57	627
K5-S7′	24	680
	28	281
	30	53
	56	320
	60	296
D5′-K1	24	306
	25	484
	45	351
	51	306
	61	328
B11-S15	10	731
	27	529
	32	140
	42	617
	49	529
B5-B10	1	214
	20	768
	33	215
	43	212
	61	73
Q2-M10	4	447
	23	444
	45	245
	46	635
	48	158
Z0′-H7	15	344
	17	973
	41	757
	52	875
Y4′-Z1	14	122
	43	555
	54	474
	62	786
T7′-P10	8	283
	11	238
	18	958
	58	283
I12′-X10’	19	152
	21	121
	36	271
	43	599
S4-O15	9	420
	13	430
	35	284
	43	253
S11-F5	20	325
	38	142
	43	470
	61	55
Z5-C4	16	213
	39	283
	59	128
E17-D14	31	254
	37	913
	50	433

Figure 1. 8-mer primer pairs for 65 human cDNAs associated with DNAREP. (A) Selected primer pairs. (B) Resulting simulated PCR products. All genes listed in Table 1 were amplified.

ing perfect matches that occur in hyper-abundant RNAs. Of the most prevalent 100 primers of 8 and 9 bases in length, about 80% and 60%, respectively, were found in the hyper-abundant RNAs and were removed.

Next, the stability of the primer template interaction needs to be considered. For successful PCR in which both primers participate, the primers need to be matched with regard to their melting temperature (T_m) on the template, especially because PCRs using such primers generally involve high-stringency conditions in order to bias sampling towards perfect matches. A+T-rich primers are unlikely to interact with the template as effectively as G+C-rich primers. We added an option to choose the range of G+C content of the primers. In the examples presented, we use a window of 50%–90% G+C content. Later, after the primer pairs to be used have been selected, the T_m of each primer can be equalized by the addition of arbitrary bases at the 5′ end. Any primers that have two or more bases of palindrome at the 3′ end are removed to avoid problems with primer dimers.

Finally, the remaining candidate oligonucleotides and their complements are combined in every possible combination and used to generate simulated PCR products from the list of sequences of interest. The number of candidate oligonucleotides used for this process can be set by the user. The default setting is 500.

The program searches for simulated PCR products in a size range set by the user. We chose a size range from 50 to 1000 bases. The upper limit of 1000 bases was chosen because products above this size are less reliable in a mixture of PCR products that also contains many more efficiently amplified products that are only a few hundred bases long. In rare instances where there is more than one simulated PCR product from a sequence using a particular primer pair, then the program retains only the shorter because this is the product preferred by PCR.

The program only searches for PCR products that have different primers at each end. There is evidence that in a competition between products that have the same primer at each end and those that have different primers at each end, those with different primers predominate after PCR. This phenomenon may be due to the formation of panhandles by the products that have the same primer at each end, giving these products a disadvantage (23). Thus, we do not wish to rely on such products to sample a particular mRNA.

The program arranges the resulting primer pairs and sequences in the list of interest into a matrix. The resulting simulated PCR products (if any) are reported in each cell of the matrix. Next, all primer pairs that share complementarity of two or more bases at the 3′ end are removed to avoid primer dimers.

The matrix is rearranged so that the primer pairs are ranked by the number of sequences in the list that they sample. The primer pair that samples the most sequences in the list is ranked first. The sequences are ranked by the number of primer pairs that generate a product. The most frequently sampled sequence is ranked first.

The matrix is simplified by including only primer pairs that sample at least a specified number of sequences. In the following results, the threshold is set so that each primer pair samples at least two sequences (or at least five). The matrix is further simplified by using an option to exclude primer pairs

Table 4: Performance of Primer Pairs Generated by the Program

List	Primers	Primer Pairs	cDNAs Sampled	Total No. Simulated PCR Products
8-mers (iterations = 5, new oligonucleotides per iteration = 500, minimum number of genes per primer pair = 2):				
DNAREP	43	22	65 (100%)	108
APO	44	22	59 (100%)	106
HNR	22	12	48 (100%)	71
GPCR	44	25	113 (100%)	179
All	108	60	285 (100%)	494
(minimum number of genes per primer pair = 5):				
All	111	66	285 (100%)	568
9-mers (iterations = 5, new oligonucleotides per iteration = 500, minimum number of genes per primer pair = 2):				
DNAREP	61	31	63 of 65 (97%)	105
APO	55	28	56 of 59 (95%)	101
HNR	48	21	48 (100%)	48
GPCR	73	41	113 (100%)	202
All	191	107	285 (100%)	490
(minimum number of genes per primer pair = 5):				
All	104	61	176 (62%)	385
10-mers (iterations = 3, new oligonucleotides per iteration = 500, minimum number of genes per primer pair = 2):				
All	191	108	228 of 285 (80%)	399

Table 5. Selectivity of Primers for the List of Interest

List	Primers	Number of Primer Pairs	Number of cDNA Samples
5A: GPCR-specific 10-mers and their complements, previously published (9), were used to generate the matrix.			
(iterations = 5, minimum number of genes per primer pair = 2):			
GPCR	20	6	18 of 113
(minus hyper-abundant)	8	0	0 of 113
5B: GPCR-specific 8-mers and their complements, derived from those previously published (9), were used to generate the matrix.			
(iterations = 5, minimum number of genes per primer pair = 2):			
GPCR	60	30	110 of 113
(minus hyper-abundant)	19	26	86 of 113

that generate products of the same size from a number of sequences in the list of interest.

The simplified matrix is used to select primer pairs. First, those primer pairs that sample sequences that are sampled by no other primer pair are selected. These primer pairs will be selected at some point, so it is most parsimonious to select them immediately. The sequences that match these primer pairs are removed from the matrix. The matrix is rearranged to show the top-ranking primer pair for the remaining sequences. This primer pair is chosen, the matching sequences are removed and so on, until the possibilities are exhausted.

When the program was tested, we found that often a few sequences were not sampled by the first matrix. To improve inclusion of sequences, the program iterates the oligonucleotide selection procedure using only those sequences that are *not* sampled by the first selected set of primer pairs or are sampled by only one primer pair. A new set of oligonucleotides that are ranked highly in this subset of previously unsampled sequences is added to the first list of primers, and a new matrix is generated and resolved. Iteration of this procedure leads to all or nearly all sequences in the list being represented. The user can set the number of iterations from 0 to 10. In the examples presented, we have used five iterations, with the 500 top new primers added to the list at each iteration.

Finally, the program generates an output of each primer pair, the sequences that each primer pair samples and the size of the product.

An example of the results of these experiments for oc-

Table 6. Partial List of Variables that Can Be Selected in the GeneUP Program

Variable	Default
Selects top-ranked oligonucleotides of any specified length, 6 bases and longer	8 bases
Oligonucleotides excluded if they occur in an “excluded” list of sequences	Human repetitive elements
Exclude oligonucleotides that match all but one base at 5′ end in the excluded list	Not excluded
Remove oligonucleotides that can easily form primer dimers	Removed
Range of G+C content	50%–90% G+C
Range of PCR product sizes	50–1000 bp
Maximum acceptable number of products in list that are the same size	2
Minimum (and maximum) number of sequences sampled by each primer pair	2 (min), 50 (max)
Perform iterations of oligonucleotide selection for sequences that are not sampled in the first matrix	2
Select the number of top oligonucleotides to be used at each iteration	500
Two statistical primers per pair or one statistical primer and one anchored primer at 3′ poly(A) tail for PCR	Two statistical primer primers
Select primer pairs from matrix starting with least-sampled sequence or most-sampled sequence that fits criteria	Start with most-frequently sampled sequence

tamers is presented in Figure 1. Table 4 summarizes data for octamers and nonamers for the four lists and for a fifth list consisting of the other four lists melded into one list of 285 cDNAs. Pairs of decamers sampled 80% of the list of 285 cDNAs using primer pairs that target at least two cDNAs. Octamers can sample all 285 cDNAs, even when the minimum number sampled per primer pair is raised to five. A complete set of data for all five lists is available (**ftp.skcc.org/mcclelland/geneupm.cl**).

The algorithm we used does not consider whether any particular cDNA is sampled by more than one primer pair in the selected list. Indeed, in all four lists, a typical cDNA was sampled by two primer pairs (Table 4, last column). This redundancy in sampling can be an advantage because it should partly ameliorate circumstances where one primer pair happens to be targeted to a region of a mRNA that is difficult to amplify.

Twenty primers of ten bases in length were developed by others for sampling GPCR cDNAs (9). These primers sampled 18 of the 113 GPCRs in our list (Table 5A). We then used every possible 8-mer within these 10-mers (a total of sixty primers) as input into GeneUP, and these were very effective, sampling 96% in the GPCR list (Table 5B). Interestingly, most of these primers had perfect matches in hyper-abundant RNAs. When these were removed, the remaining primers sampled still sampled an impressive 76% of cDNAs in the GPCR list (Table 5B). By comparing the yield of primer pairs, it is probable that GeneUP will perform as well as the Monte Carlo method (9), while allowing additional constraints to be imposed.

The rate at which primer pairs also sample other cDNAs not in the list is a matter of concern. To approximate the rate of such sampling, we used the primers selected for each list to determine if they would sample the other lists. Using a total of 151 primers on the “wrong lists” in 4 000 000 tests, we predict that each 8-mer primer pair would match perfectly about one in every 4000 mRNAs in a typical cell. So each primer pair would sample five other mRNAs in addition to the set of mRNAs of interest. A similar calculation for 8-mers derived from the 10-mers of Lopez-Nieto and Nigam (9) indicated a similar or higher rate of sampling of other mRNAs.

DISCUSSION

In this theoretical paper, we have devised a program that can select pairs of PCR primers that sample multiple sequences from a list of cDNAs. Table 6 summarizes the variables that can be set in this program. The program is capable of simulating PCR products to find primer pairs that match all or nearly all cDNAs in a list. For lists of the size we have used, 48–285 cDNAs, octamer primers give slightly better coverage than nonamers after removing primers that occur in hyper-abundant RNAs. This phenomenon may be less dramatic for very long lists where nonamers, or longer oligonucleotides, have a better chance of occurring in a sufficient number of genes.

Coverage of short lists of phylogenetically related cDNAs (HNR and GPCR) by nonamers is very effective (100% in both cases). This is not surprising. Even after removal of the many primer pairs that generate products of the same size due to PCR between conserved motifs, there still remain acceptable primer pairs that consist of one conserved motif and one statistically common primer.

The primers we have selected for the GPCR list performed favorably compared with those chosen by a different method (Compare Tables 4 and 5). For example, after removing oligonucleotides that occur in hyper-abundant RNAs, GeneUp yielded 25 primer pairs that sampled all 113 GPCR in the list. The Monte Carlo method yielded 26 surviving primer pairs that sampled 76% of the cDNAs. The effect of removing primers that occur in hyper-abundant RNAs is very dramatic, eliminating virtually all 9-mer and 10-mer primers selected by the Monte Carlo method.

The strategy we have used is designed to nearly maximize coverage, with the least number of primers. Because only a selected number of primer pairs are examined and because a heuristic method, rather than an exhaustive method, is used to pick primer pairs from the matrix, the strategy does not always pick the smallest possible set of primer pairs. However, an exhaustive strategy could be prohibitive because it would consume more computing power than is practical.

The method presented here is only one of the possible heuristic approaches to optimize selection of simulated PCR products from the matrix. Another method to select primer pairs is to start with the least-frequently sampled sequence in the matrix and select the primer pair that samples that sequence and the most other sequences in the list. The sequences that match that primer pair are removed. The matrix is rearranged, and the remaining least-sampled sequence is chosen. The top-ranked primer pair that samples that sequence is chosen. The sequences that are sampled by that primer pair are removed. The matrix is rearranged. The iteration continues until the possibilities are exhausted. This alternative has been implemented and can be chosen by the user. To date, we have found that this method generally leads to slightly less redundancy of coverage but also samples slightly fewer genes in the list (data not shown). Many other methods to pick primer pairs from the matrix can be envisioned.

One of the enduring limitations with the strategy we have outlined is the fact that oligonucleotides that are highly ranked in our list of interest are likely to be quite common in mRNAs, in general. The exclusion of hyper-abundant RNAs does not exclude the other 10 000 or more mRNAs in the cell that are not in the list of interest. We can expect that some of these other sequences will also have perfect or near-perfect matches with the primers, oriented correctly for PCR and at an appropriate distance apart. Thus, primers selected by the program increase the probability that the sequences of interest are given an opportunity to be sampled but will not necessarily exclude other sequences from being sampled in the same mixture of PCR products. It is not the intention of the program to select primers that have perfect matches exclusively with a number of the sequences of interest, while having no matches in other sequences. Rather, the purpose of the program is to ensure that the sequences of interest are among the best matches, so as to maximize the chance that these sequences will occur in the mixture of PCR products.

The other RNAs of most concern are those that are abundant in the cell because these will give a prominent PCR product. While the genes we are interested in will probably be sampled, if expressed, products from more abundant mRNAs may predominate and obscure or reduce the yield of the products of interest. We have taken the first step to minimize this problem by excluding the most abundant RNAs in the cell. However, in the future, it might be possible for this strategy to be taken a step further. Accumulating information on the relative rank of mRNAs in cells may allow the identification of the top 100 or 500 mRNAs that occur in most cell types in humans. A list of such genes would be valuable because it could be used to exclude primer pairs from consideration if, for example, more than one of these more abundant RNAs was likely to give a PCR product with a primer pair or if the expected products might be a similar size to the PCR product of interest.

Alternatively, it may be possible to improve the program to check all known Expressed Sequence Tags (ESTs) from the species and determine how often a particular pair of primers samples the whole list. Pairs that sample above a threshold could be excluded. ESTs are available in GenBank® (**http://www.ncbi.nlm.nih.gov/dbEST/index.html**).

Even with the caveat that other mRNAs will usually be sampled by any primer pairs chosen, it is worth noting that the mixture of PCR products is enriched for the intended mRNAs, even if other undesired RNAs are also sampled. Thus, these mixtures should be effective probes in differential hybridization experiments against clones for the expected mRNAs because the complexity of the probe is much lower than total cDNA (19). This could be of particular interest in strategies that are used to array clones or oligonucleotides on chips where the complexity of the probe can be a limiting factor in detecting rare transcripts. For recent reviews, see References 11 and 17.

Presently, the short primers selected using computer methods have additional arbitrary bases appended at the 5′ end when used in PCRs (9). These extra bases are used to ensure efficient PCR under the high-stringency conditions needed to select the best matches at the 3′ end. The extra bases can also be used to balance out the G+C content and thus the T_m of each primer in the primer pairs. These additional arbitrary bases naturally lead to some concern that primers will be biased towards the subset of sequences that happen to have good matches with the arbitrary extension. Thus, it is worth considering other strategies. One idea might be to use degenerate primers. This would allow some primers of 10 or more bases long to occur many times in short lists of sequences. We are currently performing the necessary biological experiments to determine if this would be a fruitful avenue for further program development.

Another variation on the method we use here may be of interest to many readers who wish to sample the 3′ ends of cDNAs (8). A single statistically selected primer can be combined with an anchored oligo(dT) primer that has one or more non-thymidine bases at the 3′ end and that primes at the border of the polyadenylation site in a large selected fraction of the mRNAs. This alternative has been implemented as an option for the program.

In this paper, we have largely ignored redundancy in the sampling of cDNAs in the list of interest while concentrating on maximizing coverage. We are currently working on a modification of the program that would sample the maximum number of mRNAs in long lists, such as all known human ESTs or all known open reading frames in a bacteria, with the least redundancy.

ACKNOWLEDGMENTS

M.M. was partially supported by a generous gift from Sidney Kimmel and by NIH Grant Nos. CA68822, NS33377 and AI34829. G.P. was partially supported by EU Grant No. BIO4-CT95-0130. T.T. was supported by a fellowship from the German DFG. A pending patent is licensed to Phenotypics Corporation.

REFERENCES

1. **Amariglio, N. and G. Rechavi.** 1993. Insertional mutagenesis by transposable elements in the mammalian genome. Environ. Mol. Mutagen. *21*:212-218.
2. **Caetano-Anolles, G., B.J. Bassam and P.M. Gresshoff.** 1991. DNA amplification fingerprinting using very short arbitrary oligonucleotide primers. Bio/Technology *9*:553-557.
3. **Carlberg, C., R. Hooft van Huijsduijnen, J.K. Staple, J.F. De Lamarter and M. Becker Andre.** 1994. RZRs, a new family of retinoid-related orphan receptors that function as both monomers and homodimers. Mol. Endocrinol. *8*:757-770.
4. **Donohue, P.J., D.K. Hsu and J.A. Winkles.** 1997. Differential display using random hexamer-primed cDNA, motif primers, and agarose gel eletrophoresis. Methods Mol. Biol. *85*:25-35.
5. **Grillo, G., M. Attimonelli, S. Liuni and G. Pesole.** 1996. CLEANUP: a fast computer program for removing redundancies from nucleotide sequence databases. Comput. Appl. Biosci. *12*:1-8.
6. **Hattori, M., S. Kuhara, O. Takenaka and Y. Sakaki.** 1986. L1 family of repetitive DNA sequences in primates may be derived from a sequence encoding a reverse transcriptase-related protein. Nature *321*:625-628.
7. **Kariya, Y., K. Kato, Y. Hayashizaki, S. Himeno, S. Tarui and K. Matsubara.** 1987. Revision of consensus sequence of human Alu repeats—a review. Gene *53*:1-10.
8. **Liang, P. and A.B. Pardee.** 1992. Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction [see comments]. Science *257*:967-971.
9. **Lopez-Nieto, C.E. and S.K. Nigam.** 1996. Selective amplification of protein-coding regions of large sets of genes using statistically designed primer sets. Nature Biotechnol. *14*:857-861.
10. **Mangelsdorf, D.J., C. Thummel, M. Beato, P. Herrlich, G. Schutz, K. Umesono, B. Blumberg, P. Kastner, M. Mark, P. Chambon et al.** 1995. The nuclear receptor superfamily: the second decade. Cell *83*:835-839.
11. **Marshall, A. and J. Hodgson.** 1998. DNA chips: an array of possibilities [In Process Citation]. Nature Biotechnol. *16*:27-31.
12. **McClelland, M., C. Petersen and J. Welsh.** 1992. Length polymorphisms in tRNA intergenic spacers detected by using the polymerase chain reaction can distinguish streptococcal strains and species. J. Clin. Microbiol. *30*:1499-1504.
13. **Mullis, K.B. and F.A. Faloona.** 1987. Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction. Methods Enzymol. *155*:335-350.
14. **Pearson, W.R., G. Robins, D.E. Wrege and T. Zhang.** 1995. A new approach to primer selection in polymerase chain reaction experiments. Ismb. *3*:285-291.
15. **Pesole, G., N. Prunella, S. Liuni, M. Attimonelli and C. Saccone.** 1992. WORDUP: an efficient algorithm for discovering statistically significant patterns in DNA sequences. Nucleic Acids Res. *20*:2871-2875.
16. **Quentin, Y.** 1994. Emergence of master sequences in families of retroposons derived from 7sl RNA. Genetica *93*:203-215.
17. **Ramsay, G.** 1998. DNA chips: state-of-the art. Nature Biotechnol. *16*:40-44.
18. **Stone, B. and W. Whorton.** 1994. Targeted RNA fingerprinting: the cloning of differentially-expressed cDNA fragments enriched for members of the zinc finger gene family. Nucleic Acids Res. *22*:2612-2618.
19. **Trenkle, T., F. Mathieu-Daude, J. Welsh and M. McClelland.** Reduced complexity probes for DNA arrays. Methods Enzymol. (In press).
20. **Welsh, J., K. Chada, S.S. Dalal, R. Cheng, D. Ralph and M. McClelland.** 1992. Arbitrarily primed PCR fingerprinting of RNA. Nucleic Acids Res. *20*:4965-4970.
21. **Welsh, J. and M. McClelland.** 1990. Fingerprinting genomes using PCR with arbitrary primers. Nucleic Acids Res. *18*:7213-7218.
22. **Welsh, J. and M. McClelland.** 1991. Genomic fingerprints produced by PCR with consensus tRNA gene primers. Nucleic Acids Res. *19*:861-866.
23. **Welsh, J. and M. McClelland.** 1991. Genomic fingerprinting using arbitrarily primed PCR and a matrix of pairwise combinations of primers. Nucleic Acids Res. *19*:5275-5279.
24. **Welsh, J. and M. McClelland.** 1992. PCR-amplified length polymorphisms in tRNA intergenic spacers for categorizing staphylococci. Mol. Microbiol. *6*:1673-1680.
25. **Welsh, J., C. Petersen and M. McClelland.** 1991. Polymorphisms generated by arbitrarily primed PCR in the mouse; application to strain identification and genetic mapping. Nucleic Acids Res. *19*:303-306.
26. **Williams, J.G., A.R. Kubelik, K.J. Livak, J.A. Rafalski and S.V. Tingey.** 1990. DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. Nucleic Acids Res. *18*:6531-6535.
27. **Woods, C.R., Jr., J. Versalovic, T. Koeuth and J.R. Lupski.** 1992. Analysis of relationships among isolates of Citrobacter diversus by using DNA fingerprints generated by repetitive sequence-based primers in the polymerase chain reaction. J. Clin. Microbiol. *30*:2921-2929.
28. **Yoshikawa, T., G.Q. Xing and S.D. Detera Wadleigh.** 1995. Detection, simultaneous display and direct sequencing of multiple nuclear hormone receptor genes using bilaterally targeted RNA fingerprinting. Biochim. Biophys. Acta *1264*:63-71.

Received 19 November 1997; accepted 18 February 1998.

Address correspondence to:

Dr. Michael McClelland
Sidney Kimmel Cancer Center
10835 Altmar Row
San Diego, CA 92121, USA
Internet:mmcclelland@skcc.org