BIOINFORMATICS

Vol. 15 no. 2 1999
Pages 93–105

A computer-driven approach to PCR-based differential screening, alternative to differential display

G. Giacomo Consalez[1], Andrea Cabibbo[1], Anna Corradi[1], Cristina Alli[1], Milena Sardella[1], Roberto Sitia[1] and Riccardo Fesce[2]

[1]Department of Biological and Technological Research (DIBIT), San Raffaele Scientific Institute (HSR), Via Olgettina 58, 20132 Milano and [2]DIBIT-HSR and Consiglio Nazionale delle Ricerche, Cytopharmacology Center, Milano, Italy

Received on January 12, 1998; revised and accepted on November 5, 1998

Abstract

Motivation: *Polymerase chain reaction (PCR)-based RNA fingerprinting is a powerful tool for the isolation of differentially expressed genes in studies of neoplasia, differentiation or development. Arbitrarily primed RNA fingerprinting is capable of targeting coding regions of genes, as opposed to differential display techniques, which target 3′ non-coding cDNA. In order to be of general use and to permit a systematic survey of differential gene expression, RNA fingerprinting has to be standardized and a number of highly efficient and selective arbitrary primers must be identified.*

Results: *We have applied a rational approach to generate a representative panel of high-efficiency oligonucleotides for RNA fingerprinting studies, which display marked affinity for coding portions of known genes and, as shown by preliminary results, of novel ones. The choice of oligonucleotides was driven by computer simulations of RNA fingerprinting reverse transcriptase (RT)-PCR experiments, performed on two custom-generated, non-redundant nucleotide databases, each containing the complete collection of deposited human or murine cDNAs. The simulation approach and experimental protocol proposed here permit the efficient isolation of coding cDNA fragments from differentially expressed genes.*

Availability: *Freely available on request from the authors.*

Contact: *fesce.riccardo@hsr.it*

Introduction

The analysis of differential gene expression focuses on molecular mechanisms involved in major biological processes, such as cell differentiation, cell division, embryonic development, neoplastic transformation and growth. A multitude of techniques have become available in recent times to isolate differentially expressed genes (see, for example, Fargnoli *et al*., 1990; Ausubel *et al*., 1995; Velculescu *et al*., 1995; Diatchenko *et al*., 1996). In 1992, Liang and Pardee first described a new, differential screening technique, based on reverse transcriptase polymerase chain reaction (RT-PCR), which they named Differential Display (DD). In this technique, cDNAs are synthesized by means of anchored oligo-dT primers, to select subsets within given mRNA populations. First-strand cDNAs are subsequently PCR amplified using the same downstream oligo-dT primer and an upstream random decamer. The complex PCR product is fractionated through a polyacrylamide gel and detected by autoradiography thanks to the incorporation of a radioactive dNTP in the PCR reaction. The technique aims at revealing bands corresponding to differentially expressed genes from a background of constitutively expressed products. Various refinements of the above protocol have been published (Bauer *et al*., 1993; Liang, 1994; Hadman *et al*., 1995; Liang and Pardee, 1995; Rohrwild *et al*., 1995; Tokuyama and Takeda, 1995; Zhao *et al*., 1995; Consalez *et al*., 1996). However, products obtained from DD gels derive almost exclusively from non-coding regions of genes, whose sequence analysis provides no cues regarding protein function.

A different RNA display protocol was developed by Welsh *et al*. (1992), to permit internally primed PCR amplification of cDNAs. In this protocol, named RAP-PCR, only arbitrary primers are used in the radioactive PCR amplification step. Unlike DD, arbitrarily primed RNA fingerprinting has not been approached systematically to maximize coverage of genes expressed in mammalian systems within a finite number of PCR amplifications. In other words, no rationally designed panel of primers is available for a fairly exhaustive survey by RAP-PCR of gene expression in a given biological system.

We have recently developed a modified RAP-PCR protocol featuring highly reproducible banding patterns, marked efficiency and sensitivity, and a non-random affinity for coding regions. This protocol has allowed several research groups to isolate differentially expressed genes from various model systems ranging from rodents to humans (Corradi *et*

© Oxford University Press

al., 1996; Malgaretti *et al*., 1997; Cabibbo *et al*., 1998; Guttinger *et al*., 1998; Covini *et al*., in press; Dragoni *et al*., 1988; Mariani *et al*., 1988; Rossetti *et al*., in preparation).

In this paper, we describe a computer analysis of human and mouse cDNA databases aimed at selecting a panel of the best primers to use in our approach. A similar, but distinct, strategy was proposed by Lopez-Nieto and Nigam (1996). The collection of reagents generated through our approach and the technical updates described here make us able to propose internally primed, PCR-based RNA fingerprinting as a reasonably simple, systematic tool for the analysis of differential gene expression, and as a workable, advantageous alternative to differential display.

Systems and methods

Simulation of mRNA PCR in nucleotide (nt) databases

PCR simulations were run on two non-redundant (nr) databases, obtained from a combination of human or mouse sequences deposited in the Genbank and EMBL nucleotide sequence databanks (accessed through the GCG Wisconsin package, Version 8.1-UNIX, August 1995) (Devereux *et al*., 1984). One arbitrary 12 nt primer sequence was used at a time, i.e. each primer was assumed to anneal in a degenerate fashion to both the sense and antisense strand.

The nr human and mouse databases were obtained by selecting human or mouse entries containing at least 1000 bp of coding sequence (CDS). In order to decrease redundancy, variable regions of immunoglobulins and T-cell receptors were eliminated, and all pairs of sequences sharing a word in their product descriptions were compared by the FASTA algorithm (Pearson, 1994); the shorter one was eliminated when >95% identical to the other. Intronic regions were eliminated from genomic sequences, thereby generating new transcribed sequence files containing uninterrupted cDNA.

Annealing of the primers was simulated by searching both strands in the nr databases for the sequence of each primer, by means of the FINDPATTERNS program in the Wisconsin GCG package, permitting a maximum of three mismatches. All pairings with one or more mismatched base(s) among the last four (at the 3′ end) were excluded as not suited to prime a PCR. A simulated PCR product was scored whenever a pairing occurred on the sense strand and, 100–1000 bp downstream, on the antisense strand.

For each primer, simulated PCR products were tagged with a CDS flag if they contained any amount of coding sequence, or with a UTR flag if they only contained a portion of 5′ or 3′ untranslated region. Each primer could be attributed an ‘efficiency’ score (total number of simulated PCR products in the sequence database) and a ‘selectivity’ score (percentage of the simulated PCR products at least partially comprised of coding sequence). All simulations were performed using a Sun Workstation; the procedures used for generating nr sequence databases and simulating RT-PCR are freely available.

A crucial aspect in assessing the validity of the approach proposed here is to exclude the possibility that differences in ‘efficiency’ observed among random primers may be fortuitous. Thus, the distribution of the number of products per primer, obtained through the simulation experiments, was compared to the one calculated for a randomly scrambled sequence database, considering that (i) all primers were constituted by eight Gs or Cs and four As or Ts, (ii) each sequence in the databank had a certain proportion of Gs or Cs and (iii) a perfect match was required for four bases at the 3′ end, whereas up to three mismatches were allowed over the first eight bases from the 5′ end of each primer. The computations of the expected product number distribution are described in the Appendix and were performed with the Matlab software (Matworks, Natik, MA) on a personal computer.

Standard molecular techniques

All standard molecular techniques employed in this paper were performed as described in *Current Protocols in Molecular Biology* (Ausubel *et al*., 1995).

RNA fingerprinting

Total RNAs were extracted in duplicate by the cesium chloride method (Sambrook *et al*., 1989), digested with 4 IU DNase, phenol–chloroform extracted, ethanol precipitated and resuspended. From 1 μg of each duplicate total RNA sample, reverse transcriptions (RT) were carried out with a $(dT)_{16}$ primer. From 10 μl of a 1:10 dilution of each RT product, radioactive PCR reactions were performed in 50 μl final volumes with single 12mer primers (final concentration 4 μM); 1.5 mM final [$MgCl_2$]; 200 μM final [dNTP]. PCR conditions were as follows: 3 min at 94°C, 2 min at 80°C when 2.5 IU Taq polymerase (Perkin Elmer) was added (hot start), followed by 35 cycles of 1 min at 94°C, 1 min at 50°C, 1 min at 72°C, with a final elongation step of 5 min at 72°C. Then, 0.1 μl [α^{32}P]dCTP (3000 Ci/mmol) or 0.2 μl [α^{33}P]dCTP (1000–3000 Ci/mmol) were added to each reaction mix. Amplification products were separated on 5% denaturing polyacrylamide gels and visualized by autoradiography. Differentially displayed bands were cut from gels and DNAs electroeluted overnight at 60 V in dialysis bags. Bands were reamplified using the same 12mer primers as in the RNA fingerprinting experiment, and cloned into the *Eco*RV site of pBluescript II SK (Stratagene), as previously described (Consalez *et al*., 1996). Clones corresponding to differentially displayed bands were selected from the background of unrelated products, as described (ibidem).

Sequence analysis of cloned products

Data bank searches (Genbank, GenEmbl, SwissProt and PIR) were run through the BlastN and BlastX network servers (Altschul *et al*., 1990). Additional sequence analysis was carried out using the GCG package (Version 9.0).

Results

Computer simulations

Computer simulations of RNA fingerprinting PCR experiments were run on custom-made human and murine nr nucleotide databases generated as described in Systems and methods.

Simulations in a human database. In a first series of simulations (MAN12.8), 10 000 12-character strings were generated at random, to represent dodecanucleotide primers, with the initial requirement that they contain eight Cs or Gs and four As or Ts, reflecting higher C and G contents in coding regions than in 3′ untranslated regions. Primers containing either stop codons (TAA, TAG, TGA) in the sense strand or ≥4 homonucleotide stretches (AAAA, CCCC, etc.) were discarded (criteria a and b, respectively). Also discarded were primers with palindromic 5′ and 3′ ends, ≥4 successive complementary bases long (criterion c). The above criteria were aimed at biasing the primers towards the CDS (a), and at enhancing the efficiency of PCR experiments (b and c).

The first 1000 primers considered acceptable according to criteria a, b and c were challenged against the human nr database (2085 sequences, 4.4 Mb total DNA, 72.8% CDS, see Systems and methods), to simulate the PCR. Primers yielding >100 but <250 simulated PCR products were considered adequate (114 primers); 345 primers were labeled as inefficient and 44 as ‘too efficient’ (primers yielding an excessive number of products might target repetitive or low-complexity templates, and result in low amplification efficiencies in the experimental phase); 497 primers were excluded because they contained >5 out of 8 bases at the 3′ end identical to a previously accepted primer (criterion d). This criterion was aimed at reducing the chance of repeatedly targeting the same sequences.

Figure 1A illustrates a histogram of simulated PCR product numbers obtained with the series of 503 accepted primers based on criteria a–d. Also illustrated is the probability density function (%) computed for the same set of primers against a randomly scrambled sequence database (see the Appendix for details on the computation of this curve). The observed distribution is non-random, featuring markedly overcrowded shoulders. This indicates the existence of significant numbers of particularly inefficient primers, as well as of particularly efficient ones, and points to the possibility of selecting a panel of valuable PCR primers based on the present ‘simulated gene fishing’ approach.

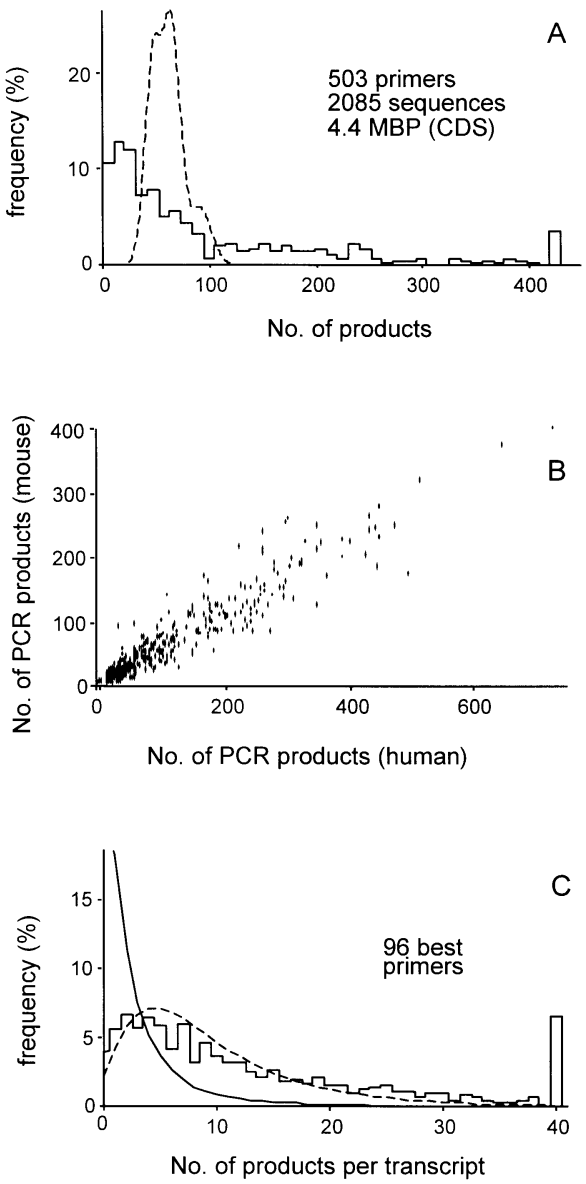


Fig. 1. (**A**) Histogram of the number of simulated CDS-containing PCR products per primer (503 primers tested on a human non-redundant cDNA database). Also shown is the expected distribution of product numbers per primer against a randomly scrambled sequence database (dashed line). Notice the large excess of poorly efficient and highly efficient primers. (**B**) Scatter plot of the number of simulated PCR products in CDS, yielded by each primer when tested on the human (*x*-axis) or mouse (*y*-axis) cDNA databases. (**C**) Analysis of exhaustivity and redundancy on the results of simulated PCR [96 ‘best’ primers (see the text) tested on a human non-redundant cDNA database]. Shown is the distribution of the number of simulated PCR products per transcript. The solid line represents the distribution of product numbers per transcript expected against a randomly scrambled sequence database. The dashed line represents the distribution expected after increasing the theoretical probability of matching by a factor equal to the ratio of observed to expected mean product numbers per transcript.

Simulations in a mouse database. The same panel of 1000 primers, generated for simulations in the human database, was tested on the mouse nr nucleotide database (MOUSE 12.8), containing 1041 sequences comprised of 2.95 Mb total nucleotide sequence (72.5% CDS), in order to check whether species-specific features in sequence composition played a major role in determining the efficiency scores of different primers. In this case, 193 'good' primers and 326 inefficient or too efficient ones were obtained; 481 primers were excluded due to criterion d. The distribution of simulated PCR product numbers per primer was qualitatively similar to the one obtained in the human database. Figure 1B is a scatter plot of the number of simulated PCR products obtained in the human (*x*-axis) versus the mouse (*y*-axis) database, with the same primers. The two sets of values correlate very well (correlation coefficient $r = 0.947$).

Degree of coverage of the primer panel. Next, we set out to address the issue of exhaustivity of a panel comprised of the 96 best primers. It may be argued that the primers selected because of their high efficiency in yielding simulated PCR products may be directed towards subpopulations of genetic sequences within the expressed sequence database. If this were the case, the number of transcripts not targeted by any primer (failures) should be higher than expected based on the mean number of simulated products per transcript. Figure 1C describes a simulation run on the human nr database and illustrates that this is not the case. In particular, it shows the number of primers targeting each transcript. The 96 most efficient primers yield more products in the actual transcribed sequence database than can be calculated for a randomly scrambled transcribed sequence database. When such an increase in efficiency is taken into account, theoretical distributions (dashed line in Figure 1C) fit the results of the simulations, arguing against a bias of our primer panel towards specific subpopulations of target sequences. Basically, our data speak for the existence of good and bad primers (non-random distribution), but fail to reveal the existence of unexpectedly good or bad templates (random distribution).

Effect of base composition of the primers. The choice of primers rich in Cs and Gs was driven by the goal of targeting coding regions. However, it became necessary to ascertain whether this unbalanced base composition in the primers could determine a bias in the choice of target sequences from the database. To address this issue, the same simulations described above were performed using 12 nt primers composed of six As or Ts and six Cs or Gs (MAN12.6, MOUSE12.6). As far as coverage, the results were superimposable on those obtained with the CG-rich primer set (not shown).

Use of degenerate primers. We reasoned that the introduction of a partially degenerate position at the 3′ end of each primer would lead to an increase in product numbers, and enhance the exhaustivity of our optimal primer collection.

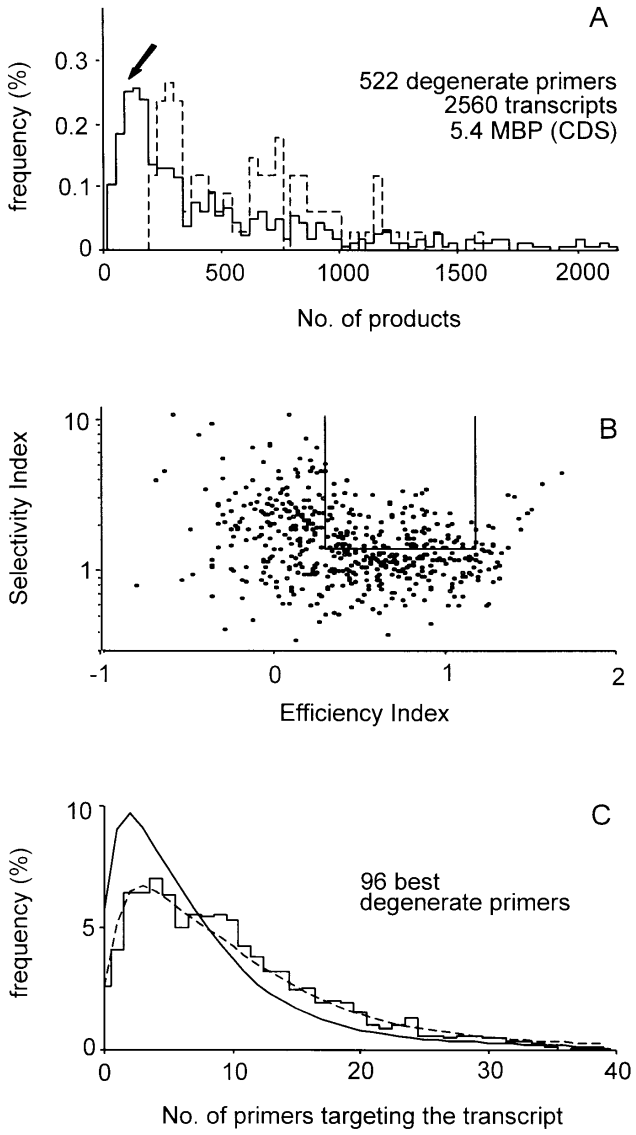


Fig. 2. Efficiency and selectivity of pairs of dodecanucleotide primers containing a partially degenerate nucleotide at the 3′ end (W or S), tested against a human non-redundant cDNA database. (**A**) Histogram of the number of simulated PCR products per primer pair (solid line: 522 primer pairs; dashed line: 96 best pairs). (**B**) Scatter plot of the selectivity index (SI) versus the efficiency index (EI). See the text and Appendix for the meaning and computation of the two indices. The area enclosed in the square corresponds to EI values between 0.3 and 1.18 (i.e. 2–15 times the modal number of products) and SIs > 1.4. Primer pairs falling in this range were included in the optimal primer set, and ordered according to their SIs. The resulting primer set is reported in Table 1. (**C**) Distribution of the number of different degenerate primers yielding simulated PCR products from each transcript. Solid and dashed lines are the expected distributions before and after correcting the theoretical probability of matching as in Figure 1C.

Table 1. Ninety-six primers selected for their high efficiency in yielding large numbers of PCR products, expressed as efficiency index (EI), and for their selectivity for coding regions versus untranslated regions of transcripts, expressed as selectivity index (SI). See the text and Appendix for the significance and computation of the two indices

ID.	sequence	S.I.	E.I.	ID.	sequence	S.I.	E.I.	ID.	sequence	S.I.	E.I.
942	ACGCCATCGACC/G	6.39	1.84	280	GCCGGGAACTTC/G	2.34	5.9	192	AGCCGGAGGATG/C	1.81	7.5
688	AAGCTGCTCGCG/C	4.99	2.07	103	ATCCTGCACCGC/G	2.33	3.28	745	AGAGTGCGCTCG/C	1.79	2.77
522	GGCACATTGCGG/C	4.5	1.96	529	AGGTACCCGTGC/G	2.31	4.98	112	CCTGCCGGAAGA/T	1.79	6.94
567	CCAGATGCCCGA/T	4.44	2.06	445	TGTTGTGGCGGC/G	2.3	3.01	952	CAGGCCACATCG/C	1.76	2.74
328	GCAGCATCCGGA/T	3.86	2.75	701	CGGCTATCGGCT/A	2.28	2.2	791	TCCTGCGGATCG/C	1.76	2.98
60	CGATCACAGCGG/C	3.63	2.28	302	GCGGACCTCATG/C	2.27	7.68	164	CAGGTCACGGAG/C	1.74	7.74
95	TCGATGCCGCTG/C	3.35	8.03	212	CTCTCCGATGCC/G	2.26	3.7	918	CCGGAAAGCACG/C	1.74	2.51
417	ATGGCAACGGCG/C	3.32	5.77	237	CCGGGTCCTCAT/A	2.24	6.79	108	CTTGGTCGTGCC/G	1.72	2.71
2	TCTGGGAACCGG/C	3.19	2.65	115	GGCGCCTACTTC/G	2.23	5.09	292	CGTGCAGTTCCC/G	1.71	3.55
187	TGCTGCAGGACC/G	3.17	7.61	674	CGGCTTTCGTGC/G	2.16	2.22	372	TCTCCGCGGTCA/T	1.71	4.27
88	CGTGGGCAACCT/A	3.13	6.9	20	ATGCCCAAGCGC/G	2.13	2.11	799	AGACCGTGGACG/C	1.7	2.33
814	GAGCTTCACCGC/G	3.13	2.32	470	CAGCAAGTCGGC/G	2.09	2.81	366	GAGGAACCGGAC/G	1.68	10.69
780	GAGGCGACGATC/G	3.1	3.92	176	AAGCCGACCTCC/G	2.08	5.79	286	GACCACCGTGTG/C	1.68	6.21
402	TCGTCGACGGTG/C	3.08	1.89	156	GCGCATCCAGCT/A	2.08	14.82	355	CCAGCGTGTTCC/G	1.68	2.12
282	ACAGGCGATGCC/G	3.08	3.06	109	GGCATTCCGCAC/G	2.04	4.32	910	CACCTACCGAGC/G	1.66	2.57
648	AGAGCAGGGCGA/T	2.93	2.01	320	GCACCGACTGGT/A	2.02	9.16	418	CCAACGAGTCCC/G	1.65	2.45
91	CTTCTCCCGGTC/G	2.8	4.82	137	ACCAAGCCCACC/G	1.99	11.2	114	CTGGTCACGTGC/G	1.64	3.88
685	TGTGGAGCCGGT/A	2.79	2.38	138	CCTGTCCGTCCT/A	1.93	18.49	224	CCCGTCTACCAC/G	1.63	12.37
151	CGGCACATCTCC/G	2.68	4.95	230	GGTCCCAATGGG/C	1.93	10.68	313	GCCCTTCGTCAG/C	1.63	5.53
333	GGCCGCATTGGA/T	2.66	6.48	357	GTCCTGCGGGTT/A	1.92	1.91	149	CACAACCGGCAG/C	1.63	10.82
97	GCAGAAGCCGTG/C	2.64	2.6	423	AGCGGAGCATCC/G	1.92	1.91	453	GTGCCACGCATC/G	1.62	2.6
29	CGGTCATGGTCG/C	2.58	2.26	283	AGCTCTGCGAGC/G	1.91	7.73	119	GCTCGATCAGGC/G	1.6	6.3
309	GGCCGAAGACCA/T	2.57	5.91	23	CCGCATGTCCAC/G	1.9	19.46	542	AGGACGCGCATC/G	1.6	3.12
850	CCAGCACTTCGC/G	2.54	2.17	560	CGCCCTGGAACT/A	1.9	3.07	580	TGGCCAACTGCG/C	1.59	3.93
468	AGCCATTCGGGC/G	2.53	2.86	868	GTCGCCGGCAAT/A	1.89	1.83	185	TCGTCACCAGCC/G	1.59	6.81
80	GTGTTGGTGCCC/G	2.47	7.53	755	GCTGCCGCCAAT/A	1.88	3.52	463	GTGGACGGTGCA/T	1.59	6.2
180	TGGACGTTGGCC/G	2.42	8.05	219	CGCTGTCGAGGA/T	1.87	5.99	132	TCGTGGCTGCAG/C	1.58	12.19
24	GGAGAAGCTGCC/G	2.39	10.91	595	AGGGCTTTCGGC/G	1.87	2.34	290	GCCTCCTGGAGT/A	1.56	9
824	AGGCGGACATCG/C	2.35	2.82	130	GGTGCTCAGCAG/C	1.85	15.4	56	TGGCTGGGATGG/C	1.54	8.4
195	AGCAGCTCGTGG/C	2.35	8.91	96	CCTTGGAAGCCC/G	1.82	8.35	332	GCGGATGCGGAA/T	1.53	4.02
52	CAATACGGGCCC/G	2.35	1.97	507	GCGGTCGAAGAC/G	1.82	4.33	163	ACGTGCCCAGCA/T	1.53	11.55
511	ACAAGGGCACGG/C	2.35	1.87	324	TCTGCCGGGTCT/A	1.82	6.8	51	TGCCGACTCTGC/G	1.51	15.7

Thus, we went back and repeated computer simulations using 12 nt CG-rich primers, containing a partially degenerate base (W or S) at their 3′ ends (representing primer pairs, rather than single primers). As expected, the use of degenerate primers in simulations gave rise to an increased product number, with respect to non-degenerate ones (Figure 2A).

Efficiency versus selectivity for coding regions. The selectivity for CDS was examined together with the efficiency in yielding high numbers of simulated PCR products. In particular, 522 degenerate primers were tested and the panel of 96 ‘best’ ones was generated by assigning an efficiency index (EI) and a selectivity index (SI) to each primer. The EI reflects the efficiency in yielding PCR products in general (i.e. the total product number), while the SI reflects the affinity of a primer for coding regions (i.e. the ratio of coding region products to the total number of products). See the Appendix for the computation of EIs and SIs. Degenerate primers with an EI between 0.3 and 1.18 (i.e. a product number 2–15 times the modal value) and an SI > 1.4 were selected (Figure 2B). The 96 degenerate primers with the highest SIs were grouped in the ‘best primers’ panel (Table 1).

Our data show that the distributions of product numbers per transcript, obtained by employing the 96 best degenerate primers, are shifted to the right with respect to expectations calculated for a database of scrambled cDNA sequences. The same applies to the number of primers targeting each transcript (Figure 2C). This evidence confirms that the 96 best primer set is particularly efficient. The expected distributions, once corrected for the efficiency ratio (dashed curve in Figure 2C, see Appendix), reasonably agrees with the observed distributions. Again, this argues against the existence of good or poor templates for our primer set; in other words, the primers fail to exhibit a bias towards a subset of genes in the database.

Coverage versus redundancy. A series of simulations with different parameters were run to investigate the exhaustivity and redundancy of the approach proposed here. Simulations were run on either the human or the murine nr cDNA databases. Selected panels of 12 base primers with either six or eight Gs

or Cs out of 12 were employed. In some simulations, the whole set of primers was tested, and in some a maximum of only two mismatches was allowed in the simulated annealing process. Figure 3 illustrates the relationship observed between the mean number of products per transcript and the number of failures (transcripts not targeted by any primer) in nine simulations. It can be noticed that the number of no hits (lack of exhaustivity) decreases as the mean number of products per transcript (redundancy) increases. However, the decrease is much slower than expected based on simple Poisson statistics, and 90% coverage is reached at the cost of a redundancy factor of 10. The failure rates are in agreement with the expectations computed for scrambled databases when all primers are used (empty triangles), while they decrease significantly when high-efficiency primers are employed.

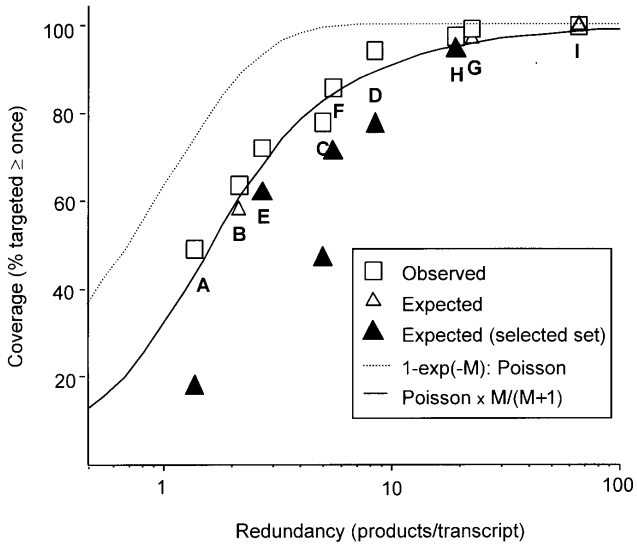


Fig. 3. Relationship between redundancy (average number of simulated PCR products per transcript) and coverage (percentage of transcripts targeted by at least one primer). Simulations performed on human (Hu) and mouse (Mo) non-redundant sequence databases. All simulations with 12 nt primers; three mismatches allowed, unless otherwise specified. Squares: observed coverage versus redundancy in various simulations. (**A**, **B**) Hu, degenerate primers (maximum two mismatches, eight Cs or Gs), best 96 and all primers, respectively; (**C**) Mo, non-degenerate primers (six Cs or Gs), 57 primers yielding >100 products; (**D**) Mo, non-degenerate primers (eight Cs or Gs), 93 primers yielding >100 products; (**E**) Hu, non-degenerate primers (six Cs or Gs), 68 primers yielding >100 products; (**F**, **G**) Hu, non-degenerate primers (eight Cs or Gs), 96 best primers and all primers, respectively; (**H**, **I**) Hu, degenerate primers (eight Cs or Gs), 96 best and all primers, respectively. Triangles: corresponding expected coverage, calculated for scrambled sequence databases; the observed numbers of failures are lower than expected in simulations performed with sets of selected primers (filled symbols). The dashed line represents the predictions of Poisson statistics. The continuous line is an arbitrary analytical function which suggests the following empirical relationship:

$$\text{coverage} = \frac{\text{redundancy}}{\text{redundancy} + 1} \cdot 100 \cdot [1 - \exp(-\text{redundancy})]\%$$

Experimental assessment of the primer panel

As a result of the elaboration described above, 25 primers were synthesized and tested at the bench (12 non-degenerate, 13 degenerate), to assess the correspondence between theoretical predictions and experimental results. As templates, we used total RNA from HepG2 cells cultured in different redox conditions (Cabibbo *et al.*, 1998), and from embryonic and postnatal mouse brain territories (Corradi *et al.*, in preparation; Alli *et al.*, in progress).

Set up of optimal amplification conditions. The optimal annealing temperature for our primer panel was determined by testing different annealing temperatures in a range between 45 and 53°C. The best results were obtained at 50°C. Lower temperatures caused weaker, fuzzier banding patterns, whereas no bands were seen with 52 or 53°C in the annealing step. The optimal number of degenerate nucleotides was also determined experimentally, by testing primers containing no degenerate position, one degenerate position at the 3′ end nucleotide, or two degenerate positions at nt 11 and 12. The best and most reproducible results were repeatedly obtained with one site of partial degeneration (W or S) at nt 12, or the 3′ terminal nucleotide.

Annealing of the primers to the target sequences. As a further step, we analyzed the mode of annealing in our experimental conditions with respect to the annealing constraints set in our simulations. Clones obtained by RNA fingerprinting were sequenced and aligned to nucleotide databases. Products identical to deposited cDNAs or ESTs were analyzed to determine where exactly in the transcript the primer had annealed, with how many mismatches, and where the mismatches had occurred in the primer sequence. As shown in Figure 4A, the experimental annealing conditions represent an excellent approximation of the simulated ones. In only two cases out of 29 did a mismatch occur at any site within the last four bases at the 3′ end of the oligonucleotide. As it turns out, in all other cases the last four bases of the primer and template matched perfectly, defining a 3′ 4-base stretch critical for annealing and elongation when using a 12mer in our PCR conditions. Throughout the remaining length of the primer, annealing could occur even in the presence of up to four mismatches.

Correlation of theoretical and actual PCR product numbers. What can be inferred from the analysis of our simulations is not the total number of products to be expected in each experiment. This value is affected by a cohort of experimental

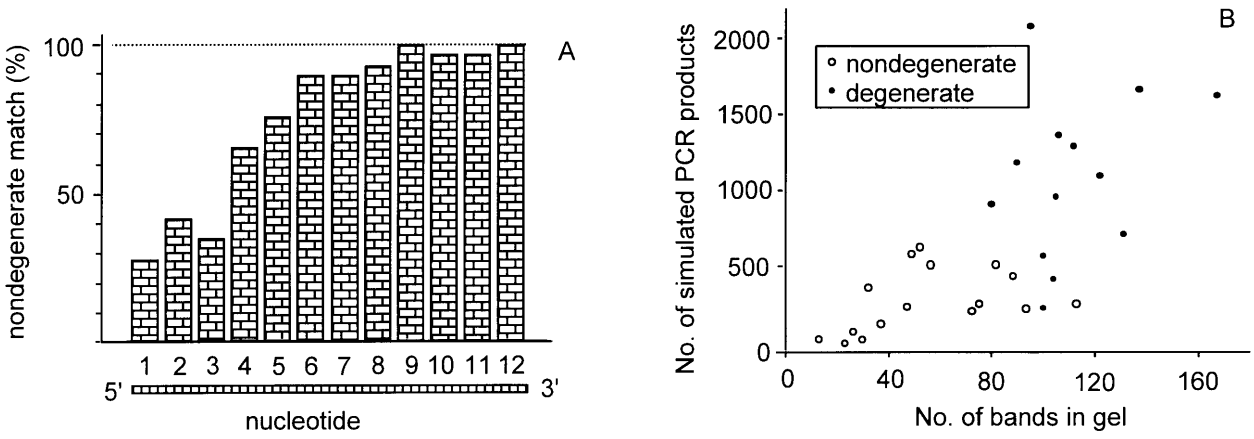


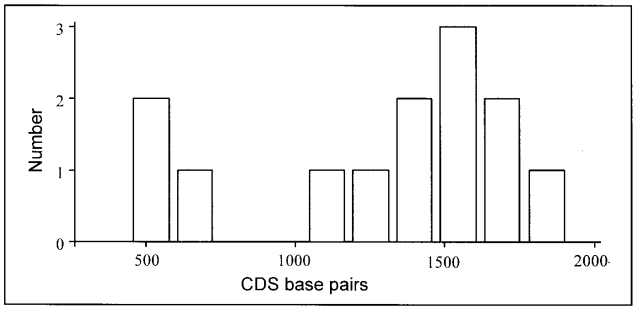
Fig. 4. (**A**) Diagram illustrating the mode of annealing of 12mer primers to known sequences deposited in nucleotide databases (GenBank, DBEST). Shown is the percentage of perfect matches, as opposed to degenerate annealing, observed in 29 cases employing eight different primers. *y*-axis*:* percentages of non-degenerate annealing measured at each residue of the 29 primers analyzed. (**B**) Correlation between product numbers predicted by simulations and numbers of bands observed in RNA fingerprinting experiments performed with the corresponding primers. Simulations were performed on mouse non-redundant cDNA database, using 12 nt primers (eight C-G, four A-T), degenerate and non-degenerate. Thirteen primers were arbitrarily chosen among those yielding low, medium and high numbers of simulated PCR products; PCR experiments were performed as described in Systems and methods, and only clearly discernible bands were recorded. Examples of RNA fingerprinting gels obtained with our method are published elsewhere (Malgaretti *et al*., 1997; Cabibbo *et al*., 1998).

variables, including the complexity of the gene pool in a given tissue and the stringency of experimental conditions. In fact, one goal of the present study was to establish general criteria useful to predict the relative efficiency of each computer-generated oligonucleotide sequence based on the frequency of its occurrence in transcribed and translated regions. To this end, we proceeded to count the numbers of bands observed in RNA fingerprinting gels performed with some of the primers generated as described above. An extremely conservative approach was adopted in counting bands: only clearly distinct products evident after overnight exposure were taken into account. Our data suggest that the use of the radioactive tracer [α^{32}P]dCTP, as opposed to [α^{33}P]dCTP, results in a significant underestimate of product numbers, giving rise to fuzzy, overlapping bands. The mean number of bands obtained by using each of 12 non-degenerate primers was 66; the mean number of bands obtained by using each of 13 degenerate primers was 111, a significant increase in efficiency. While on a purely statistical basis the introduction of one partially degenerate nucleotide is expected to yield an ~4-fold increase in product numbers in our system, a lower increase was observed experimentally, probably due to limiting factors in the PCR reaction. When our data were arranged in the form of a scattergram (Figure 4B), a positive correlation could be observed between theoretical expectations and relative experimental efficiencies. In particular, the correlation indicates that primers expected to be efficient always give rise to a high number of products (absence of false positives), whereas some of the primers expected to work poorly do in fact perform well in the experimental phase (presence of false negatives). Extrapolating these preliminary data, it can be inferred that a differential screening study employing the best 96 primers in our collection should permit a survey of >10 000 bands. An estimate of the coverage provided by this figure is dependent upon the degree of redundancy and the complexity of the gene pool in each given tissue or cell line. At this stage, an experimental assessment of redundancy is of limited significance, due to the small number of products analyzed so far.

Selectivity of primer panel for coding sequences. Thirty-three 'differentially displayed' bands obtained by RT-PCR with each of nine different 12mers were cloned and sequenced, 29 of which displayed open reading frames (ORFs) throughout their lengths. Of these, 20 represented known CDS or their homospecific/cross-specific homologues; nine additional new ORFs were analyzed with the GRAIL program (Roberts, 1991), which predicted an excellent coding probability in six out of nine cases. The relatively low number of cloned and sequenced products makes it impossible to correlate expected and observed ratios of coding to non-coding cDNAs, and, in fact, the ratio of coding to non-coding cDNAs reflects that of translated to untranslated RNA sequence in mammalians. However, it may certainly be inferred from these preliminary results that the internal primer panel described here enhances access to coding re-

Table 2. Transcript sizes (kb) and coding sequence (CDS) sizes (bp) of cDNAs deposited into the Genbank database, identified by PCR-based differential screening using our computer-assisted gene fishing approach. A histogram of ORF lengths illustrates their distribution

gene name	mRNA (kb)	CDS (bp)	Ref. or GB accession #
Csrp1	1.8	582	D88793
Mab21	3.0	1080	AF040945
O/E3	3.8	1728	U92703
Eva	3.4	648	AF030454
HARP1	2.8	1200	Rossetti et al., in preparation
Ezrin	2.7	1761	X60671
EDF1	1.0	444	Dragoni et al., 1998
ZFM/SF1	3.5	1914	Covini et al., in press
KIF3C	3.8	1587	AF018164
Cytokeratin 17	1.6	1362	Z19574
Thioredoxin reductase	3.4	1494	AJ001050
cERB-A	2.8	1371	X04707
PLAP	2.8	1593	M12551



gions by two orders of magnitude when compared to the differential display approach.

Size of amplified cDNAs. Because the described amplification strategy is based on the occurrence of two copies (one sense and one antisense) of a degenerate 12mer within a single mature transcript, we set out to determine whether our approach would lead to the preferential amplification of large size cDNAs, corresponding to large transcripts. Table 2 describes the sizes of transcripts and ORFs from 13 published cDNAs identified through our approach. These transcripts range in size between 1.0 and 3.8 kb. The corresponding ORFs range in size between 582 and 1914 nt. These preliminary figures suggest that small transcripts are amplified as efficiently as larger ones, likely due to (i) high levels of degeneracy in the primer annealing step and (ii) high-frequency occurrence of specific sequence motifs in coding regions.

Discussion

The basic idea of computer-assisted gene fishing, illustrated here, is to select, among random-sequence PCR primers, the most efficient and selective ones, as judged from the results of computer simulations of PCR procedures on non-redundant cDNA sequence databases. The simulation approach represents a forceful oversimplification of experimental conditions and is governed by rigid rules: in the case described here, hybridization was assumed to occur whenever a perfect 4-base match at the 3′ end of the primer occurred, with no more than three mismatches in the remaining portion, and a relevant PCR product was assumed to occur whenever a pairing occurred on the sense strand and a second pairing occurred on the antisense strand, 100–1000 bp downstream. These assumptions overlook the occurrence of (i) highly degenerate annealing on particularly abundant mRNA species, which can yield detectable bands in PCR gels; (ii) hairpins generating gaps in primer–template doublets; (iii) limiting PCR reagent concentrations which may disfavor mRNA species present at low copy numbers. Still, the results reported here support the value of simulation as a rational approach to designing efficient and selective primers for RNA fingerprinting.

The notion of employing 12 bp PCR primers selected according to criteria other than pure chance is supported by several lines of evidence. Experimentally, large differences are observed in the number of PCR products generated from the same cDNAs by ‘high-efficiency’ versus ‘low-efficiency’ primers. It is well known that the frequency of occurrence of nucleotide ‘words’ of a given length varies widely among different types of DNA sequences (introns, coding regions, etc.) (Claverie *et al.*, 1990). This suggests that specific 12 nt sequences, when used as PCR primers, may be endowed with widely different efficiencies and possibly selectivities for coding versus non-coding regions. Indeed, all simulations indicate that the distributions of the number of simulated PCR products per primer depart markedly from the ones computed for a randomly scrambled database (probability density function). A large excess of particularly efficient and particularly ‘poor’ primers was observed in all simulations, which confirms the notion that the proposed approach might help in selecting particularly favorable 12 nt primer sequences for RNA fingerprinting. When the significance of these findings is assessed by challenging the same set of primers against the human and murine sequence databases, the numbers of PCR products obtained in the two simulations correlate very well, indicating that the unexpectedly high or low efficiencies of some primers do not arise from aberrations in the composition of the databases used, but rather from intrinsic differences in efficiency among primers. In other words, this confirms that some ‘genetic words’ have particularly high or low probabilities of occurring in CDS (Claverie *et al.*, 1990), with no major differences between phylogenetically related genomes.

The predictivity of simulation in selecting efficient primers appears to be confirmed by the analysis of the number of PCR products obtained experimentally with a reduced panel of primers. The actual probability of hybridization under the

experimental conditions described was obviously unknown, as were the exact sizes of the cDNA pools analyzed. The absolute numbers of PCR products were lower than in the simulations; however, the relative efficiencies of the tested primers were in reasonable agreement with our predictions.

A series of criteria were adopted in selecting the random-sequence primers to be tested by PCR simulation. These criteria were aimed at excluding primers which would likely generate technical problems (e.g. those containing homonucleotide stretches), at biasing the primers towards coding regions (a fixed ratio of eight Cs or Gs to four As or Ts was used, and primers containing stop codons in the sense strand were excluded) and at obtaining exhaustivity (at least four out of eight nucleotides at the 3′ end were unique for each selected primer). The last constraint turned out to be quite restrictive, in that after selecting about 100 primers, many further random sequences had to be generated in order to find a new, compatible one. Conceivably, this may lead to an enhanced coverage by our primer panel.

As reported above, most of the PCR products cloned so far using the primers in the panel contain significant ORFs, either throughout their whole lengths, or at one end. This supports the idea that a panel of primers selected as described here should make it possible to address coding regions in mRNA in a majority of cases, permitting a prediction as to the nature of corresponding peptide sequences, and the establishment of cross-specific relationships with deposited sequences from distant phyla. Although an increasing number of transcribed sequences are deposited in nucleotide databases, catalogs are only complete for one model eukaryotic organism, and a long time may elapse before full-length cDNAs become available for a significant number of organisms. Likewise, in the case of large genomic sequences deposited in the framework of organism-specific genome sequencing projects, the availability of a coding sequence tag would circumvent the need for the lengthy and error-prone process of identifying and splicing transcribed sequences at the computer. Thus, the prompt recognition of a newly discovered sequence as a member of a phylogenetically conserved gene family represents a significant advantage of our approach over previously described ones. The results described here are more relevant when one considers that, in its current version, our protocol utilizes oligo-dT-primed first-strand cDNA, thus shifting the cDNA pool towards 3′ ends. An experimental update in our approach will require first-strand cDNA synthesis using internal primers, rather than oligo-dT primers, to enhance targeting of CDS further. To date, this has not been done lest a significant portion of products might contain rRNA sequences, and is the current subject of methodological work by our group.

An additional question to address was whether the primers selected because of their high efficiency in yielding simulated PCR products were directed towards subpopulations of genetic sequences. Three criteria were considered: (i) the relationship between the length of the sequences and the number of products they yielded showed no sign of clustering into subpopulations (not shown); (ii) the distributions of the number of simulated PCR products per sequence never displayed multiple modes and were generally smooth and continuous; (iii) although such distributions were shifted towards high numbers of products (especially when obtained with the sets of 96 best-performing primers) with respect to the expectations for scrambled sequence databases (Appendix), they were well fit by simply increasing the matching probability by a factor equal to the ratio (observed/expected) of mean numbers of products; the latter approach also yielded adequate predictions of the percentage of non-targeted transcripts ('silent sequences'). All this argues against the existence of subpopulations of nucleotide sequences with significantly different probabilities of being recognized by the sets of primers employed here.

A more difficult problem to tackle is redundancy, i.e. the production of many amplification products from each transcript in the mRNA library used. A balance between redundancy and exhaustivity must be reached empirically. In our simulations, the percentage of 'silent sequences' was closely approximated by the reciprocal of the mean number of products from each sequence (a measure of redundancy). Our data suggest that, as the number of products increases, the proportion of 'silent sequences' decreases more slowly than expected from Poisson statistics, where the fraction of failures is predicted by exp(–mean). The departure from Poisson behavior, observed in the simulations, is predicted theoretically, based on the heterogeneous probability of being sampled for sequences with dissimilar length and composition. An experiment employing 96 primers with some 100 bands per primer would yield ~10 000 discernible PCR products; assuming that a typical cell expresses some 20 000 genes, this would correspond to an average hitting rate of 0.5 PCR products per expressed gene. According to our theoretical predictions and simulations, under these conditions coverage should approximate 13% (as opposed to the theoretical upper limit of 39% for a purely random, Poisson distribution). Thus, the method proposed here is certainly not aimed at obtaining exhaustive coverage of differential gene expression. Much higher numbers of primers (and PCR products) would be needed; for example, in order to fish out 90% of the 20 000 expressed genes, one should analyze ~200 000 bands. Classical differential display with similar numbers of PCR products should obtain similar coverage rates, since available data suggest that the probability of yielding PCR products by differential display is also markedly heterogeneous (genes expressed in high number of copies have a much higher probability of yielding PCR products; Bertioli *et al*., 1995; our unpublished data). Thus, the main differences should not regard coverage, but bias at low

coverage levels: whereas differential display tends to produce PCR products for more abundantly expressed genes, the approach proposed here appears not to be biased in this regard and to produce a high percentage of PCR products in coding regions.

An approach similar to ours has been taken by others (Lopez-Nieto and Nigam, 1996). Those authors proposed and described a protocol employing each one of 30 computer-generated arbitrary 8 nt primers selected for their probability of occurrence in sense strands of coding sequences, to be used in combination with each of the reverse complement series, i.e. 29 primers. In the present paper, we propose the utilization of single, partially degenerate primers selected for their frequent and balanced appearance both in sense and antisense strands. The protocol proposed by Lopez-Nieto and Nigam entails the use of $30 \times 29 = 870$ PCR amplifications, as opposed to 96 in our schema. Although 870 primer combinations will provide greater coverage of the genes expressed in a given tissue sample than just 96 PCRs, from our analysis of the relationship between exhaustivity and redundancy, one would suspect that the gain may not be worth the effort (2–3 times as many expectedly targeted genes). More in general, we believe that one should probably not embark on a PCR-based differential screening project with the goal in mind of producing a complete catalog of differentially expressed genes; instead, a more sensible scope would be to generate a number of genetic tags helpful in initiating the dissection of functional pathways in development or differentiation active in one's system of choice. As a matter of fact, these pathways typically involve, in addition to differentially expressed genes, ubiquitously expressed genes as well as post-transcriptional and post-translational regulatory events. Other valuable, albeit more demanding, approaches are available to those researchers who wish to generate complete catalogs, rather than obtaining a sample of differential gene expression in their biological systems (Kato, 1995). More importantly, microchip-based technology will become more generally available in the medium term for high-throughput, genome science style studies in many organisms.

At difference with Lopez-Nieto and Nigam's study (1996), which described computer analysis for a large set of random primers, but focused on the experimental validation of a primer set specifically designed to target a group of G protein-coupled receptor genes, the present study has identified primers not biased towards any specific gene family, which have been utilized experimentally by a number of groups to target protein coding regions in general, and all appear to work equally efficiently in the standard experimental conditions described in the present paper (Corradi *et al.*, 1996; Malgaretti *et al.*, 1997; Cabibbo *et al.*, 1998). Furthermore, the octamers need 5′ adaptors to increase product numbers and high annealing temperatures (54°C) to decrease the background of low-efficiency/truncated amplification products (Lopez-Nieto and Nigam, 1996). However, this procedure might somewhat favor the recursive amplification of sequences partially or completely homologous to the adaptor. Our protocol employs longer primers, at a lower annealing temperature (50°C), resulting in good control of the background level, and high amplification efficiencies.

The evidence presented here demonstrates that, in our scheme, only four residues at the 3′ end of our primers need to anneal with a perfect match to their templates for the PCR to take place, while up to four mismatches are well tolerated over the remaining eight residues. Thus, 12mer primers such as the ones proposed in the present paper produce large numbers of products in our experimental conditions by virtue of partially degenerate annealing. Furthermore, as primer sequences used in experiments are identical to those used in simulations (no adaptors), effective lessons on how to refine the simulation strategy can be learned from the analysis of experimental results.

In summary, the present approach, by combining the predictive power of computer-based database analysis with the establishment of robust, repeatable experimental conditions, proposes PCR-based RNA fingerprinting as a rejuvenated, efficient approach to the analysis of differential gene expression.

Acknowledgements

We thank Carol L.Stayton and Nicoletta Malgaretti for their essential contribution to the described protocols. This methodological work was made possible by grants from the Italian Telethon (B14) and the Associazione Italiana Sclerosi Multipla (AISM) to G.G.C. and CNR target project on biotechnology to A.C. Partly supported by the G. Armenise-Harvard Foundation.

Appendix

Computation of efficiency index (EI) and selectivity index (SI)

A modal value was computed from smooth lines fit to histograms of the number of simulated PCR products per primer (see, e.g., the solid lines in Figures 1A and 2A). The efficiency index, EI, for each primer was computed as the decimal logarithm of the ratio: (number of simulated PCR products for this primer)/(modal value for all the primers tested).

The total numbers of coding and non-coding (3′ untranslated) base pairs in the non-redundant pseudo-cDNA database were counted and the ratio (total coding)/(total UTR) was computed. For each primer, the simulated PCR products were grouped as 'CDS' or '3′-UTR'. The selectivity index, SI, was computed for each primer as the ratio:

SI = [(CDS products)/(UTR products)]/[(total coding)/(total UTR)]

Computation of the expected distribution of PCR products per primer

Let the databank, **D**, be a set of N sequences, $\mathbf{D} = \{S_s | s \in [1,N]\}$. Given that S_s is a sequence composed of a_s C/G nucleotides and b_s A/T nucleotides, the probability of a G or C nucleotide in the primer matching an arbitrary nucleotide in the sequence is $\frac{a_s}{2(a_s + b_s)}$, and the corresponding probability for A or T is $\frac{b_s}{2(a_s + b_s)}$.

In order to obtain hybridization, a certain degree of matching must be obtained; here we arbitrarily decided that hybridization would occur for at least nine matching bases out of 12, with no mismatches within the last four bases at the 3′ end. Under these conditions, the probability of hybridization for a given template sequence and a given primer is a function of the fraction of C/G nucleotides in the sequence, $F_s = a_s/(a_s + b_s)$, the number of C/G in the first eight bases of the primer, n_1, and the number of C/G in the last four bases at the 3′ end, n_2.

For any specific alignment of the primer on the template S_s, we have:

$$\mathbf{p}[\textit{base (j)} \text{ of the primer matches} \mid F_s] = \begin{Bmatrix} F_s/2 & \textit{base}(j) \in \{C,G\} \\ (1-F_s)/2 & \textit{base}(j) \in \{A,T\} \end{Bmatrix}$$

$A_s = \mathbf{p}[\text{at least five out of the first eight bases match} \mid F_s, n_1] =$

$$= \sum_{j=1}^{N_1} \mathbf{p}[j \text{ matches out of } n_1 \text{ (C}\vee\text{G)}] \cdot$$

$$\mathbf{p}[\text{at least } 5-j \text{ matches out of } 8-n_1 \text{ (A}\vee\text{T)}] =$$

$$= \sum_{j=1}^{n_1} \left[C_{n_1}^{j} (F_s/2)^j (1-F_s/2)^{n_1-j} \cdot \sum_{k=5-j}^{8-n_1} C_{8-n_1}^{k} \left(\frac{1-F_s}{2}\right)^k \left(1-\frac{1-F_s}{2}\right)^{8-n_1-k} \right]$$

$$B_s = \mathbf{p}[\text{all four bases at the 3′ end match} \mid n_2] = \left(\frac{F_s}{2}\right)^{n_2} \cdot \left(\frac{1-F_s}{2}\right)^{4-n_2}$$

so that the probability of hybridization for any specific alignment of the primer on the template is $P_s = A_s \times B_s$. The average value of F_s was 0.53 (± 0.082) and in general P_s was ~$1–2 \times 10^{-4}$, its value increasing for primers with increasing numbers of C/G nucleotides in the last four positions.

Assuming that PCR products of interest would have a length, L, comprised between L_0 = 100 and L_1 = 1000 bp, then the number of combinations of two acceptable positions on a template sequence S_s, of length M_s, is:

$$C_s = \sum_{j=1}^{M_s-L_1} (L_1-L_0+1) + \sum_{j=M_s-L_1+1, j>0}^{M_s-L_0+1} (M_s-L_0-j+1) = (M_s-L_1)\cdot(L_1-L_0+1) + \sum_{k=0}^{L_1-L_0} k =$$

$$= \begin{Bmatrix} (M_s-L_0)(M_s-L_0+1)/2 \approx (M_s-100)^2/2 & \text{for } M_s < L_1 \\ (L_1-L_0+1)\cdot\left(M_s-\frac{(L_1+L_0)}{2}\right) = (M_s-550)\cdot 901 & \text{for } M_s \geq L_1 \end{Bmatrix}$$

This gives rise to a binomial distribution of the number of PCR products obtained from template sequence S_s of length M_s and a primer with given values of n_1 and n_2. Such distribution is defined by the binomial parameters $\mathbf{p} = P_s^2$ and $\mathbf{n} = C_s$; the corresponding probability density function (p.d.f.) is: $p_s(x) = C_{C_s}^{2x} \cdot P_s^{2x} \cdot (1 - P_s)^{C_s - x}$

Actually, to estimate the probability of obtaining simulated PCR products (neglecting the technical aspects connected to experimentally obtaining a PCR amplification product), the unwanted possibility of a further hybridization in between the two valid positions must be excluded. For a product of length L, this possibility has an approximate probability of $(1 - (1 - P_s)^{2L}) \approx 2L \cdot P_s$, and therefore ~$900 \times P_s$ for the average product length of 450 bp. For the usual magnitude of P_s, this factor amounts to ~10^{-2} and can be neglected.

The distributions of PCR products from the N sequences in the database $\{p_{s(x)} \mid S_S \in \mathbf{D}\}$ are expected to be independent. Therefore, the corresponding characteristic functions (ch.f.) can be computed $\{\varphi_s(u) \mid S_s \in \mathbf{D}\}$ and the ch.f. for the whole databank will simply be: $\varphi_D(u) = \exp\left(\sum_{S_s \in D} \log[\varphi_s(u)]\right)$.

From $\varphi_{\mathbf{D}}(u)$, the expected p.d.f. of the number of PCR products from the whole databank, $p_{\mathbf{D}}(x)$, is computed for each primer. Averaging over the set of primers yields the expected distribution of the number of PCR products per primer ($\boldsymbol{P_1}$). Notice that $p_{\mathbf{D}}(x)$ is necessarily equal for primers having the same values of n_1, n_2 and P_s. Thus, $\boldsymbol{P_1}$ may be multimodal (up to five peaks for n_2 = 0–4).

The same procedure is used to compute the expected p.d.f. of the number of PCR products from each sequence, $\boldsymbol{P_2}$ (in this case, the ch.f. is computed by summing the logs of the single ch.f. values over the set of primers for the same sequence).

A third distribution of interest is that of the number of 'successful' primers per sequence (i.e. yielding at least one PCR product from the sequence), $\boldsymbol{P_3}$. This is computed in the same way using the modified p.d.f., π_s, such that: $\begin{Bmatrix} \pi_s(0) = p_s(0) \\ \pi_s(1) = \sum_{j=[1,\infty]} p_s(j) \end{Bmatrix}$.

Distribution $\boldsymbol{P_1}$ is used to check whether the observed distribution of PCR products per primer departs significantly from the expectation: if a marked excess of particularly 'good' and 'poor' primers is found, this argues against a

purely random distribution of nucleotides in the sequences of the databank.

Distributions P_2 and P_3 yield information on the exhaustivity of the approach, i.e. the capability of picking out as many different sequences as possible. In particular, the shape of the p.d.f. P_3 can be compared to the corresponding distribution, obtained by the simulation experiments, to check whether any bias is present towards a subpopulation of sequences (i.e. whether some sequences are significantly more subject to amplification than others).

This approach cannot be straightforwardly applied to the distributions obtained using sets of particularly efficient primers. These distributions are obviously shifted to the right, with respect to the p.d.f. P_3, which is computed based on a purely random nucleotide composition of the databank. The size of the shift is conveniently represented by the ratio of the mean values. If the shift simply reflects an increased hybridization probability with no bias towards sequence subpopulations, the shape of the curve will be easily reproduced by computing the logarithm of the characteristic function of the expected probability, multiplying it by the ratio of the means and computing the resulting probability distribution (this is performed by applying the direct and inverse fast Fourier transforms). A reasonable agreement with the observed distribution will argue against biases in favor of specific sequence subpopulations.

References

Altschul,S., Gish,W., Miller,W., Myers,E. and Lipman,D. (1990) Basic Local Alignment Search Tool. *J. Mol. Biol*., **215**, 403–415.

Ausubel,F.M., Brent,R., Kingstone,R.E., Moore,D.D., Smith,J.A. and Struhl,K. (1995) *Current Protocols in Molecular Biology*. John Wiley and Sons, New York.

Bauer,D., Muller,H., Reich,J., Riedel,H., Ahrenkiel,V., Warthoe,P. and Strauss,M. (1993) Identification of differentially expressed mRNA species by an improved display technique (DDRT-PCR). *Nucleic Acids Res*., **21**, 4272–4280.

Bertioli,D.J., Schlichter,U.H., Adams,M.J., Burrows,P.R., Steinbiss,H.H. and Antoniw,J.F. (1995) An analysis of differential display shows a strong bias towards high copy number mRNAs. *Nucleic Acids Res*., **23**, 4520–4523.

Cabibbo,A., Consalez,G.G., Sardella,M., Sitia,R. and Rubartelli,A. (1998) Changes in gene expression during growth arrest of HepG2 hepatoma cells induced by reducing agents or TGFbeta1. *Oncogene*, **16**, 2935–2944.

Claverie,J.M., Sauvaget,I. and Bougueleret,L. (1990) K-tuple frequency analysis: from intron/exon discrimination to T-cell epitope mapping. *Methods Enzymol*., **183**, 237–252.

Consalez,G.G., Corradi,A., Ciarmatori,S., Bossolasco,M., Malgaretti,N. and Stayton,C.L. (1996) A new method to screen clones from differential display experiments prior to RNA studies. *Trends Genet*., **12**, 455–456.

Corradi,A., Croci,L., Stayton,C., Gulisano,M., Boncinelli,E. and Consalez,G.G. (1996) cDNA sequence, map and expression of the murine homolog of *GTBP*, a DNA mismatch repair gene. *Genomics*, **36**, 288–295.

Covini,N., Tamburin,M., Consalez,G., Salvati,P. and Benatti,L. Induction of ZFM1/SF1 mRNA in rat and gerbil brain after global ischemia. In press.

Devereux,J., Haeberli,P. and Smithies,O. (1984) A comprehensive set of sequence analysis programs for the VAX. *Nucleic Acids Res*., **12**, 387–395.

Diatchenko,L. *et al*. (1996) Suppression subtractive hybridization: a method for generating differentially regulated or tissue-specific cDNA probes and libraries. *Proc. Natl Acad. Sci. USA*, **93**, 6025–6030.

Dragoni,I., Mariotti,M., Consalez,G.G., Soria,M.R. and Maier,A.M. (1998) EDF-1, a novel gene product involved in human endothelial cell differentiation. *J. Biol. Chem*., **273,** 31119–31124..

Fargnoli,J., Holbrook,N.J. and Fornace,A.J.,Jr (1990) Low-ratio hybridization subtraction. *Anal. Biochem*., **187**, 364–373.

Guttinger,M., Sutti,F., Panigada,M., Porcellini,S., Merati,B., Mariani,M., Teesalu,T., Consalez,G.G. and Grassi,F. (1998) EVA, a novel member of the immunoglobulin superfamily expressed in embryonic epithelia with a potential role as homotypic adhesion molecule in thymus histogenesis. *J. Cell Biol*., **141**, 1061–1071.

Hadman,M., Adam,B.L., Wright,G.L.,Jr and Bos,T.J. (1995) Modifications to the differential display technique reduce background and increase sensitivity. *Anal. Biochem*., **226**, 383–386.

Kato,K. (1995) Description of the entire mRNA population by a 3 end cDNA fragment generated by class IIS restriction enzymes. *Nucleic Acids Res*., **23**, 3685–3690.

Liang,P. (1994) Differential display using one-base anchored oligo-dT primers. *Nucleic Acids Res*., **22**, 5763–5764.

Liang,P. and Pardee,A. (1992) Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. *Science*, **257**, 967–971.

Liang,P. and Pardee,A.B. (1995) Recent advances in differential display. *Curr. Opin. Immunol*., **7**, 274–280.

Lopez-Nieto,C. and Nigam,S. (1996) Selective amplification of protein-coding regions of large sets of genes using statistically designed primer sets. *Nature Biotechnol*., **14**, 857–861.

Malgaretti,N., Pozzoli,O., Bosetti,A., Corradi,A., Ciarmatori,S., Bianchi,M., Martinez,S. and Consalez,G.G. (1997) *Mmot1*, a new helix-loop-helix transcription factor gene displaying a sharp anteroposterior expression boundary in the embryonic mouse brain. *J. Biol. Chem*., **272**, 17632–17639.

Mariani,M., Corradi,A., Baldessari,D., Pozzoli,O., Fesce,R., Martinez,S., Boncinelli,E. and Consalez,G.G. (1998) *Mab21*, the mouse homolog of a *C. elegans* homeotic regulator, participates in cerebellar, midbrain and eye development. *Mech. Develop.*, **79,** 131–135.

Pearson,W.R. (1994) Using the FASTA program to search protein and DNA sequence databases. *Methods Mol. Biol*., **25**, 365–389.

Roberts,L. (1991) GRAIL seeks out genes buried in DNA sequence [news]. *Science*, **254**, 805.

Rohrwild,M., Alpan,R.S., Liang,P. and Pardee,A.B. (1995) Inosine-containing primers for mRNA differential display. *Trends Genet*., **11**, 300.

Rossetti,G., Impagnatiello,M.A., Orecchia,S., Bianchi,E., Croci,L., Consalez,G.G. and Pardi,R. HARP-1, a relative of Brainiac, encodes a novel intercellular contact-regulated protein and is preferentially expressed in tissues of ectodermal origin. In preparation.

Sambrook,J., Fritsch,E.F. and Maniatis,T. (1989) *Molecular Cloning: A Laboratory Manual*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Tokuyama,Y. and Takeda,J. (1995) Use of 33P-labeled primer increases the sensitivity and specificity of mRNA differential display. *Biotechniques*, **18**, 424–425.

Velculescu,V.E., Zhang,L., Vogelstein,B. and Kinzler,K.W. (1995) Serial analysis of gene expression. *Science*, **270**, 484–487.

Welsh,J., Chada,K., Dalal,S., Cheng,R., Ralph,D. and McClelland,M. (1992) Arbitrarily primed PCR fingerprinting of RNA. *Nucleic Acids Res*., **20**, 4965–4970.

Zhao,S., Ooi,S.L. and Pardee,A.B. (1995) New primer strategy improves precision of differential display. *Biotechniques*, **18**, 842–6, 848, 850.