A kinetic model for subtractive hybridization

Joel J. Milner*, Edi Cecchini and Peter J. Dominy

Plant Molecular Science Group, Institute of Biomedical and Life Sciences, Division of Biochemistry and Molecular Biology, Botany Building, Glasgow University, Glasgow G12 8QQ, UK

Received May 17, 1994; Revised and Accepted November 28, 1994

ABSTRACT

Nucleic acid sequences that differ in abundance between two populations (target sequences) can be cloned by multiple rounds of subtractive hybridization and amplification by PCR. These sequences can be cDNAs representing up-regulated mRNAs, or genomic DNAs from deletion mutants. We have derived an equation that describes the recovery of such sequences, and have used this to simulate the outcome of up to 10 rounds of subtractive hybridization and PCR amplification. When the model was tested by comparing its predictions with the published results from genomic and cDNA subtractions, the predictions of the model were generally in good agreement with the published data. We have modelled the outcomes of genomic subtractions, for a variety of genomes, and have used it to compare various strategies for enriching targets. The model predicts that for genomes of less than $5 \times 10^8$ bp, deletions of as small as 1 kbp should represent $>99\%$ of the DNA after three to six rounds of hybridization (depending on the enrichment procedure). As genomes increase in size, the kinetics of hybridization become an important limiting factor. However, even for genomes as large as $3 \times 10^9$ bp, it should be possible to isolate deletions of 5 kbp using the appropriate conditions. These simulations suggest that such methods offer a realistic alternative to chromosome walking for identifying genomic deletions for which there are known phenotypes, thereby considerably reducing time and effort. For cDNA subtractive hybridization, the model predicts that after six rounds of hybridization, sequences that do not differ in abundance between the tester and driver populations (the background) will represent $<1\%$ of the subtracted population, and even quite modestly up-regulated cDNAs should be successfully enriched. Where several up-regulated cDNAs are present, the predicted final representation is dependent on both the initial abundance and the degree of up-regulation.

INTRODUCTION

If two populations of nucleic acids are hybridized under conditions in which one, the driver, is in excess over the other, the tester, the rate of re-annealing will be determined by, inter alia, the concentration of the driver [10]. Differences between the driver and tester, in the relative abundance of particular sequences, will affect the rate of re-annealing, and can be exploited as a means of isolating sequences (the target sequences) that are either present only in the tester, or which are relatively more abundant in the tester than in the driver. This technique is known as subtractive hybridization. The driver nucleic acid (which is depleted in the target sequences), and the tester nucleic acid (which is enriched in the target sequences), are sometimes referred to respectively as the (−) and (+) populations.

Subtractive hybridization was first used as early as 1966 by Bautz and Reilly to purify phage T4 mRNA [1]. More recently a number of groups have employed variations of the technique, both to clone cDNAs derived from mRNAs that undergo up- or down-regulation (cDNA subtraction) [2–4], and to identify genomic deletions (genomic subtraction) [5,7–9]. Early techniques for cDNA subtraction generally involved one or two rounds of hybridization and used (−) mRNA to drive hybridization to (+) cDNA tester [6]. Compared with differential screening of conventional cDNA libraries, subtractive cDNA cloning offers a number of advantages, notably greater sensitivity and improved ease of interpretation. However, the technique is not without difficulties. The preparation of (−) mRNA in large amounts is not always a practical proposition; consequently, for less abundant sequences, the concentration of driver (and hence the Cq) is likely to be too low to drive hybridization to completion. The degree of enrichment is limited by the driver:tester ratio, and a single round of hybridization will only enrich adequately those up-regulated messages that are rare in the (−) population but highly abundant in the (+) population. Sequences that are only moderately abundant even after up-regulation, or which are up-regulated only to a limited extent, will still be obscured by a background of common sequences. Furthermore, the amount of cDNA remaining after hybridization can be tiny; the problem of cloning successfully such minute quantities of cDNA is not trivial.

The advent of PCR has allowed the development of improved techniques for the construction of subtractive cDNA and genomic libraries. Duguid and Dinauer [3] have published a technique that involves the ligation of linkers to the cDNA molecules. This allows PCR amplification of the tester cDNA between hybridization cycles, and multiple rounds of hybridization can be carried out. These authors used the technique successfully to clone up-regulated sequences from scrapie-infected sheep brain. We

* To whom correspondence should be addressed
have used a modified and improved technique to identify mRNAs up-regulated in dedifferentiating *Nicotiana glauca* [4].

Several laboratories have used an analogous approach for genomic subtraction [5,7,8]. Sun *et al.* have cloned the GAI locus in *Arabidopsis* by using an excess of DNA from a mutant with a 5 kb deletion in the locus as the driver and DNA from the wild-type as the tester [5]. Lisitsyn *et al.* [9] have combined genomic subtraction with kinetic enrichment and PCR amplification of a fraction of the genome (genomic sampling), adding to the potential power of the technique. The ability to identify genes by subtraction can, in theory, be applied to any organism in which deletion mutations give rise to selectable phenotypes. The development of rapid and reliable experimental protocols for the routine identification of deleted sequences that give rise to identifiable phenotypes could replace the labour-intensive method of 'chromosome walking', and is clearly of great interest.

Despite the potential of the technique, we are not aware of any rigorous theoretical analysis of either genomic or cDNA subtraction. A plethora of experimental procedures have been published, often based on intuition. This is surprising, since the kinetics of hybridization have been well understood for many years, and where both driver and tester DNAs are double-stranded, renaturation should be expected to follow simple second order kinetics [10]. We have extended these formulations to define the recovery of target sequences after hybridization and subtraction. We have used these algorithms to build a model which simulates the predicted enrichment, subtractive and kinetic, of target sequences for up to 10 rounds of hybridization and subtraction, and where data are available, we have compared these predictions with experimental observations. The model allows us to predict the most favourable conditions for genomic and cDNA subtractions.

**MATERIALS AND METHODS**

**Derivation of the model**

The model assumes that there are two populations of DNA (either genomic DNA or double-stranded cDNA), the (+) and the (−) populations. The populations are similar in composition except for a few sequences, the target sequences, which are either absent completely from the (−) population, or whose relative abundance in the (−) population is lower than in the (+) population. The object of the exercise is to enrich the target sequence. A large excess of (−) DNA, which acts as driver, is re-annealed with (+) DNA, which acts as the tester. Various methods can then be applied to the isolation of DNA enriched in the target sequence. If the driver DNA is labelled with biotin, biotinylated DNA (including driver–tester hybrids) can be removed following hybridization, using streptavidin [6]. Alternatively specific PCR primers can be used, that allow amplification of re-annealed tester–tester duplexes only [9].

For hybridization between driver and tester, we can consider the DNA as a population of individual sequences. Each may be present in one or both of the populations and will hybridize independently. If biotinylation is used as the enrichment method, for any sequence *a* (where *a* is an individual sequence or a class of sequences), the concentration of tester sequence after removal of the biotinylated DNA, is given by:

\[
C^+_\text{residual} = C^+_a + B
\]

where \(C^+_a\) is the concentration of the tester DNA sequence [i.e. derived from the (+) DNA population] which has remained single-stranded after hybridization, and \(B\) is the concentration of tester–tester duplex.

If PCR is used to selectively amplify tester–tester hybrids [9], equation (1) reduces to

\[
C^+_\text{residual} = B
\]

After completion of re-annealing, for sequence *a*, the ratio of hybridized to unhybridized molecules derived from the tester will be the same as that from the driver population provided: (i) biotinylation does not significantly affect the rate of renaturation; and (ii) other factors that affect the rate constant (such as sequence length) are the same for both DNA populations, \([10,11]\). This fraction is given by the second order rate equation \([10]\):

\[
\frac{C^+_w}{C^+_0} = \frac{1}{1 + kC^+_0 t}
\]

Since \(C^+_w\) is used to derive the tester DNA population, the total initial concentration of sequence *a* (derived from both tester and DNA populations), and the total concentration of sequence *a* (derived from both driver and tester populations) which has remained single stranded after t s. \(k\) is the apparent second order rate constant for the sequence (in mol/s/l).

The concentration of double-stranded sequence *a* in which both strands of the duplex are derived from the tester DNA is given by:

\[
B = C^+_a \frac{C^+_w}{C^+_0}
\]

where \(C^+_a\) is the concentration of the tester DNA-derived sequence *a* in all (tester–driver and tester–tester) double-stranded molecules. Equation (6) is usually expressed in the form

\[
B = C^+_a \frac{C^+_w}{C^+_0}, \text{ since when } C^+_0 \gg C^+_a, \frac{C^+_w}{C^+_0} \equiv C^+_w \equiv C^+_0. \text{ However, although this is true for the common background sequences, it is not true in the case of the target sequences and therefore we use the more general form of the equation.}

Again, at any given time, the fraction of tester DNA-derived sequence *a* which is double-stranded \((C^+_+/C^+_+)\) is equal to the fraction of total sequence *a* which is double-stranded \((C^+_w/C^+_w)\) provided that both the driver and tester DNA strands renature with the same second order rate constant. This can be derived from equation (3), the second order rate equation:

\[
\frac{C^+_w}{C^+_0} = \frac{kC^+_0 t}{1 + kC^+_0 t}
\]
Since \( \frac{C^*_{st}}{C_0} = \frac{C^*_{ds}}{C^*_{ds}} \) \( \tag{8} \)

\[ C^*_{st} = C_0^* \frac{kC^*_{ds}t}{1 + kC^*_{ds}t} \] \( \tag{9} \)

where \( C^*_{ds} \) is the total concentration (tester–tester, tester–driver and driver–driver duplexes) of sequence \( a \) which is double-stranded.

Therefore from (6) and (9),

\[ B = \frac{k(C^*_{st})^2t}{1 + kC^*_{ds}t} \] \( \tag{10} \)

Equation (2) then allows us to solve \( C^*_{residual} \) for selective PCR amplification.

For selection using biotinylated DNA, from equations (1), (5) and (10),

\[ C^*_{residual} = C_0^* \left( \frac{1 + kC^*_{st}t}{1 + kC^*_{ds}t} \right) \] \( \tag{11} \)

Equation (10) or (11) allows us to calculate the recovery of unbound sequence \( a \) following hybridization and subtraction, provided we know the initial concentrations of sequence \( a \) in the driver and tester DNAs and the apparent second order rate constant. The latter is given by:

\[ k = \frac{k_i}{G} \] \( \tag{12} \)

where \( k_i \) is the basic nucleation constant and \( G \) is the complexity in nucleotides [11].

We can calculate \( k_i \) to be \( 1.18 \times 10^6 \text{ bp mol}^{-1} \text{s}^{-1} \) for 450 bp fragments of DNA in 0.12 M sodium phosphate by using the published rate constant of 0.25 mol s^{-1} for \( E. coli \) DNA (\( G = 4.7 \times 10^6 \)) under these conditions [10,11]. Equation (12) can be corrected for hybridization with DNA fragments greater or smaller than 450 bp, and in \( Na^+ \) concentrations other than 0.18 M [11,12], and the rate constant is given by:

\[ k = \frac{Zk_iL}{450} \sqrt{450} \] \( \tag{13} \)

Where \( L \) is the average length of the smaller of the two fragments in bp [11] and \( Z \) is an empirical factor that corrects for the effect of \( Na^+ \) concentration on the rate of renaturation, and which has been estimated by Britten et al. [12].

Equations (10) or (11) plus (13) allow us to model the behaviour of a complex population of DNA sequences in a subtractive hybridization, since we may assume that the population comprises a finite number of independently re-annealing individual sequences or classes of sequence, each with its own complexity and apparent second order rate constant.

The model

Simulations of the enrichment of target sequences were run using these equations and Microsoft Excel (v. 4.0 for Windows). For simplicity, sequences whose abundance was the same in the driver and tester populations (the common background) were modelled as abundance classes. For genomic DNA, these correspond to the various classes of repetitive and single copy DNA [10], and the input variables are the proportion of the genome and copy number. In the case of cDNA subtractive hybridizations, these represent different abundance classes of mRNA [13], and the input variables are the number of mRNAs per cell and the number of mRNA species in the class. The target sequences (those which are relatively more abundant in the tester population than in the driver population) were entered individually or as classes. The model calculates \( k \) and \( C^*_{residual} \) from equations (13) and either (10) or (11) for up to 10 rounds of hybridization from input data and treats the following as variables: initial concentrations of tester and driver DNAs; the correction factor \( Z \) for the \( Na^+ \) concentration [12]; the physical length of the DNA strands; and, for each sequence or sequence class, the complexity and abundance within the tester and driver populations. Where the residual DNA is amplified by PCR, we assume that all residual sequences are amplified with equal efficiency. We assume that the amount of amplified DNA added to the hybridization reaction is sufficient to give a concentration equal to that at the first round of hybridization [4,9]. Thus after each round of PCR amplification, the input concentrations of tester and driver DNAs used in the subsequent round of hybridization are the same as the concentrations before the first round of hybridization. The efficiency of amplification is therefore not a variable in the model, but could be incorporated if the empirical relationships were known.

Copies of the model can be obtained by sending a 3.5 inch disk (DOS format) to the authors, or from the software archives via anonymous ftp (for details contact the authors by E-mail on j.milner@udcf.gla.ac.uk).

Construction and analysis of subtractive cDNA libraries

The construction of conventional and subtractive cDNA libraries has been described [4]. mRNA abundance was estimated by slot-blot hybridization.

RESULTS

Several reports of genomic subtraction involving multiple rounds of hybridization have been published [5,7,8]. These provide useful data with which to test the model, since the differences between the two DNA populations are known, either because the driver DNA contained a single defined deletion, or because the tester DNA was deliberately spiked with a known quantity of foreign DNA. We have attempted to model the genomic subtraction experiments of Sun et al. [5] (Arabidopsis DNA spiked with adenovirus DNA), and Wieland et al. [8] (human genomic DNA spiked with \( \lambda \) DNA). Using the figures shown in Table 1 for the haploid size, and the copy number and complexity of DNA sequences in the Arabidopsis and human genomes [14,15], our model predicts target enrichments of 930-fold and 500- to 1500-fold respectively, compared to the published values of 500- and 700-fold respectively. For human, but not Arabidopsis DNA, the predicted levels of enrichment of the target were highly dependent on those input variables which affect the value of \( k \), e.g. \( L^5 \) [11], some of which were difficult to estimate from the published data. Given the lack of detail (yields after each round of hybridization, etc.) in the published data, the predicted enrichments are consistent with the published values.
Table 1.

<table>
<thead>
<tr>
<th>Organism</th>
<th>1C (bp)</th>
<th>Sequence classes</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Arabidopsis</em></td>
<td>1.0 x 10^8</td>
<td>0.30 (193); 0.70 (1).</td>
<td>Leitwiler et al. [14]</td>
</tr>
<tr>
<td><em>D.melanogaster</em></td>
<td>1.5 x 10^8</td>
<td>0.15 (60); 0.85 (1).</td>
<td>Laird and McCarthy [16]</td>
</tr>
<tr>
<td><em>B.mori</em></td>
<td>5.1 x 10^6</td>
<td>0.24 (50000); 0.21 (500); 0.55 (1).</td>
<td>Gage [17]</td>
</tr>
<tr>
<td>Soybean</td>
<td>1.3 x 10^9</td>
<td>0.26 (3850); 0.24 (153); 0.43 (1).</td>
<td>Gurley et al. [18]</td>
</tr>
<tr>
<td><em>N.tabacum</em></td>
<td>2.5 x 10^9</td>
<td>0.08 (12400); 0.67 (250); 0.25 (1).</td>
<td>Zimmern and Goldberg [19]</td>
</tr>
<tr>
<td>Human</td>
<td>3.1 x 10^9</td>
<td>0.08 (300000); 0.10 (60000); 0.03 (600); 0.70 (1).</td>
<td>Britten and Kohne [15]</td>
</tr>
</tbody>
</table>

Haploid genome size (1C) and sequence class data used as input data for the model for the organisms shown. Data are based on renaturation kinetics. Sequence class data show the fraction of the genome for each class followed by, in parentheses, the mean number of copies per genome of the sequences within the class.

*Excludes the snapback/highly repetitive sequence class which re-annales at very low Cfg. Consequently, the sum of fractions of genome for all classes add up to less than 1.0.

Our prediction of the enrichment of a 5 kbp deletion in the Arabidopsis genome (99% representation after five rounds of hybridization) is not consistent with the published value of 5% [5]. This discrepancy could not be ascribed to errors in estimating the value of k. Because Sun et al. ligated the linkers after the final round of hybridization, only double-stranded DNA would have been amplified. However, correcting for this had no significant effect on predicted yield. Sun et al. reported that the biotinylation of their driver DNA was relatively inefficient. Possibly, this might contribute to the low recovery of target. We can model this by substituting U.C_0 for C_0 in equation (11), where U is the proportion of fragments that are unbiotinylated and C_0 is the initial concentration of driver. We can then determine the amount of unbiotinylated driver remaining unbound after hybridization. For U = 0.001, the recovery of the GAl target falls to 11% of the total DNA, and U = 0.002, the model predicts that if U ≥ 0.001, the representation of the target sequence reaches a maximum after three rounds of hybridization and does not increase with additional rounds. Sun et al. noted that the enrichment reached a maximum after three rounds of hybridization.

Some subtraction protocols include a PCR amplification step after each or each alternate hybridization [3,4,9] and we can use our model to test the effect of doing so. We use the acronym SHAP for subtractive hybridization, amplification by PCR. The model predicts that for values of U up to 0.01, PCR amplification carried out either after each hybridization, or after each alternate hybridization, abolishes the effect of unbiotinylated driver fragments on target recovery.

Lisitsyn et al. [9] have described the use of representational difference analysis, a kinetic enrichment approach utilizing selective PCR amplification of tester–tester hybrids (here we use the acronym SPAT, for selective PCR amplification of tester duplexes, to distinguish the method from SHAP). They further improved enrichment by combining SPAT with ‘genomic sampling’, the generation, using infrequently cutting restriction endonucleases, of a subset of sequences that are subsequently preferentially amplified by PCR. We have modelled these workers’ enrichment of phage λ and adenovirus DNA from a background of human DNA using equation (10). Rather than attempting to model directly the effect of fragment size on PCR amplification efficiency, we have approximated the effect by assuming that a subset of the genome, the sampled fraction, is efficiently amplified by PCR, the remaining sequences being incapable of amplification. This reduces the effective genomic complexity to 1C × (proportion of the genome represented by the sample fraction). Our predicted recoveries, making the arbitrary assumption that BamHI or BglII will generate sample fractions of about 0.05–0.1, fall in the region 92–96% after three rounds of hybridization. Although Lisitsyn et al. do not quantitate their recoveries, these predictions are consistent with their data. Interestingly, the predicted recovery without genomic sampling is still 27% after three rounds and 97% after four rounds of hybridization.

Cloning genomic deletions

We can compare the efficiency of enrichment using selective PCR amplification of tester–tester hybrids (SPAT), with that of SHAP (biotin selection). Figure 1 shows the predicted representation, with successive rounds of hybridization, of a 5 kbp target in the Arabidopsis and human genomes. For the purposes of comparison, we have used as input parameters for both models the hybridization conditions of Lisitsyn et al. [9], which are significantly more kinetically favourable than those of Sun et al. [5]. For SHAP we have assumed a PCR amplification after each alternate hybridization [4]; this obviates potential difficulties caused by very low concentrations of tester, whilst minimizing any possible effects of differential amplification. For Arabidopsis, the efficiencies of enrichment were similar (Figures 1a and 1b). However, for the human genome, even without genomic sampling, SPAT gives a greater enrichment of the target. The model also predicts that when biotin selection is used, repetitive DNA sequences disappear from the background first, whereas with selective PCR amplification, the low copy number DNA is removed most efficiently.

We can use the model to simulate the effects of genome size and target (deletion) size on genomic subtractions. Since real genomes vary both in size and in the proportions and copy number of repetitive DNA sequences, we have modelled initially a series of subtractive hybridizations with a set of hypothetical genomes based on the characteristics of Arabidopsis (viz. all have 70% single copy DNA; the remaining 30% comprises repetitive sequences with a copy number of 193 per haploid genome [14]). This test the effect of genome size alone. Our model presupposes either SHAP with a PCR amplification step after each alternate hybridization [equation (11)] or SPAT [equation (10)]. We have
used the input variables from Sun et al. [5] and assumed that genomic sampling has not been carried out. Figure 2a shows the predicted recoveries using SHAP of 1 and 5 kbp target sequences with genomes ranging in haploid genome size (1C) from $1 \times 10^8$ (Arabidopsis) to $1 \times 10^{10}$ bp (100 times Arabidopsis). Figure 2b shows the same data modelled using SPAT. Figures 2c and 2d show the predicted recoveries, using SHAP and SPAT respectively, of target sequence whose size comprises a fixed proportion ($5 \times 10^{-5}$) of the corresponding genome, equivalent to a 5 kbp deletion in Arabidopsis.

We have modeled the outcome for the Drosophila melanogaster, Bombus mori, soybean, Nicotiana tabacum (tobacco) and human genomes. For simplicity, each genome was assumed to be composed of a limited number of sequence classes corresponding to the various classes of repetitive and low copy number DNA, as determined from the kinetics of renaturation. Table 1 shows 1C, and the proportions and copy numbers of the various classes of sequence, taken from the kinetics of renaturation [15–19]. When the predicted recoveries for real and hypothetical genomes of the same size are compared (see Figures 2a–d) the differences are small. Predicted recoveries of target sequences are affected to a limited extent only by variations in the proportions of single copy and repetitive DNA.

For organisms with genomes of $1C \leq 2 \times 10^8$, e.g. Arabidopsis or Drosophila, cloning deletions as small as 1 kbp should be easy, target sequences representing >99% of the DNA after four cycles of hybridization, and little difference is apparent between enrichments under SHAP and SPAT. As 1C increases from $5 \times 10^8$ to $5 \times 10^9$, the kinetics of hybridization become increasingly limiting, and SPAT is more efficient than SHAP, even in the absence of genomic sampling.

To enhance the kinetics of hybridization, increased hybridization times, higher driver concentrations, greater driver:tester ratios, longer DNA fragments, and the use of techniques that enhance the rate of reassociation, e.g. phenol emulsion reassociation technique (PERT) [20] or solvent exclusion [21], may be effective. At a driver DNA concentration of 3.125 mg/ml the effective enhancement under PERT is only 2.2-fold [12,20]. However, compared to the rate in 1.0 M NaCl, the relative enhancement in the presence of 11% dextran sulphate and 1.5 M NaCl is 11.9-fold [12,21], although driver DNA concentration is limited to 1.0 mg/ml, reducing the achievable enhancement [12,21] to about 4-fold. Figure 3 shows the predicted effect for the N.tabacum and human genomes of increasing target size under standard conditions [5] and enhanced conditions (11% dextran sulphate, 1.5 M NaCl, 500 bp driver DNA). Under the standard
conditions, the predicted recovery of even very large targets is limited, particularly for SHAP. However, with enhanced conditions, recoveries of small deletions (1–5 kbp) should be feasible using either SHAP or SPAT.

Cloning up-regulated mRNA by cDNA subtraction

As with genomic subtraction, we have tested the model using published data, although only a limited amount of suitable and complete data was available in the literature. We have previously reported the subtractive cDNA cloning of two mRNAs that are up-regulated in \textit{N. glauca} following transfer to callus-inducing medium (CIM) [4]. In order to provide precise data with which to test the model, we have measured, using slot-blot hybridization, the levels of the \textit{edeA} and \textit{edeB} mRNAs in \textit{Nicotiana} pith before and 8 h after transfer to CIM. Using the level of hybridization to an actin probe as a standard, taking the level of actin mRNA to be 0.01% of the total transcript level [2] and assuming 225 000 mRNA molecules per tobacco cell [21], we estimate the levels of \textit{edeA} mRNA to be 15 and 90 molecules/cell and the levels of \textit{edeB} mRNA to be 100 and 220 molecules/cell in somatic tissue and 8 h after transfer to CIM respectively (blots not shown). We have carefully re-assessed the abundance of \textit{edeA} and \textit{edeB} recombinants in the subtractive library [4], and taking into account the background of non-recombinants, we now estimate that recombinants that hybridize to \textit{edeA} and \textit{edeB} represent 46 and 4% respectively of the library.

Goldberg et al. [22] have reported that \textit{N. tabacum} leaf mRNAs fall into three abundance classes (for details see the legend to Figure 4). These data have been used to model the levels of the
mRNAs that do not alter in abundance following transfer to CIM. For our initial test of the model, we have assumed that edeA and edeB are the only up-regulated mRNAs. Using the conditions of hybridization of Cecchini et al. [4], enrichments using SHAP are shown in Figure 4a. Common background cDNAs derived from highly and moderately abundant mRNAs are rapidly depleted and fall to less than 1% of the total population after two cycles of hybridization. After five rounds of hybridization, common sequences represent less than 1% of the total cDNA, and edeA and edeB cDNAs reach a steady-state level of 94 and 5% respectively. On the basis of hybridization kinetics, Goldberg et al. estimated the number of molecules per cell of each low abundance mRNA to be 17. However, this figure represents an average for what is in reality a rather heterogeneous class. We have simulated variation in the abundance of the various rare mRNAs by arbitrarily splitting the class into two, a low abundance class (5850 mRNAs, 28 copies/cell) and a very low abundance class (5850 mRNAs, 5 copies/cell). Under these circumstances the
predicted final steady-state levels of *edeA* and *edeB* cDNAs are virtually unaltered (Figure 4b). However, seven rounds of hybridization are now required to reduce the levels of very low abundance cDNAs to below 1%. *edeA* to *edeB* ratios predicted by the model (20:1) are thus consistent with the observed ratios (11.5:1).

We do not know the identities of any mRNAs other than *edeA* and *edeB* that are up-regulated, although the presence of at least two other cDNA clones in the subtractive library (our unpublished observations) suggests that such mRNAs exist. We have modelled several hypothetical scenarios that postulate additional up-regulated mRNAs at different concentrations in the (+) and (−) populations. Figure 4c shows the effect of an additional class of six species, each present at 5 and 25 copies/cell in the (−) and (+) populations respectively (5-fold up-regulation). Figure 4d shows the effect of a single additional cDNA present at 20 and 200

---

**Figure 3.** Effect of increasing target size on recovery of targets in genomic subtractions using tobacco and human DNAs. Symbols are as shown in the figure. Standard (std) conditions (solid symbols) are as shown in the legend to Figure 2. Input parameters for enhanced (Enh) conditions (open symbols) are: driver DNA concentration = 1.0 mg/ml, tester DNA concentration = 0.1 mg/ml, Na\(^+\) concentration = 1.5 M, dextran sulphate concentration = 11% (w/v), Z = 91 (from [12,21]), hybridization time per cycle = 20 h, fragment length = 500 bp. (a) Data modelled using SHAP; (b) data modelled using SPAT.
copies/cell in the (–) and (+) populations respectively. Introducing such extra cDNAs reduces the final levels of both edeA and edeB. The extent of reduction depends on the levels and numbers of additional classes of cDNAs in the (+) and (–) populations. Provided these additional cDNAs are of moderate abundance or show only limited up-regulation, e.g., as in Figure 4c, the final levels of edeA relative to edeB remain in the range 19:1–25:1. However if additional species are both abundant and highly
up-regulated (e.g. Figure 4d), the effect on the levels of edeB is greater than on the levels of edeA. The presence of extra cDNAs has almost no effect on the number of rounds of hybridization required for levels of enrichment to reach a steady-state. This is almost exclusively a function of the input parameters for the low abundance common mRNAs. Under a variety of scenarios, the predicted ratio of edeA to edeB cDNA clones (−20:1 under conditions where additional up-regulated species are of limited abundance) remain in good agreement with the observed ratio within the subtractive library.

These data suggest that the model can be useful in predicting the outcome of subtractive cDNA cloning for a variety of scenarios. With up-regulated mRNAs at various concentrations and a background population similar in abundance to that in a tobacco leaf cell, six–seven rounds of hybridization are generally sufficient to enrich the up-regulated cDNAs to more than 99% of the population (data not shown). Assuming an extra very low abundance class as described above increases this to seven rounds of hybridization. High levels of enrichment can be achieved even if the up-regulated mRNA is present at low abundance, e.g. a single mRNA species at 5 and 20 copies respectively in the (−) and (+) populations (Figure 4e). Where more than one up-regulated sequence is present, and if all show equal relative levels of up-regulation, their abundance relative to each other remains similar, although not identical, after subtraction (not shown).

However, where several species show differential up-regulation, the situation is complex and the final relative abundance depends on both concentration and extent of up-regulation (cf. edeA and edeB Figures 4b–d).

Removal of the less abundant common background cDNAs is kinetically limited, and highly abundant common cDNAs are depleted before rare ones. Consequently, self-subtraction, in which the (+) and (−) cDNA populations are the same, will result in enrichment of the least abundant sequences (Figure 4f).

Self-subtraction may therefore prove helpful for cloning rare mRNAs, since three or four rounds of hybridization will reduce the background level of cDNAs from abundant messages.

Although the SPAT technique was originally designed for genomic subtraction, it should also be applicable to cDNA subtraction, and we have modelled the same scenarios using equation (10) (data not shown). Recoveries of targets are generally similar to those predicted using SHAP, e.g. under the scenario shown in Figure 4b, under SPAT, six rounds of hybridization are required to reach predicted levels of 89% and 11% for edeA and edeB. However, under SPAT, low abundance common cDNAs are depleted more efficiently than highly abundant ones. Consequently any contaminants in the subtractive library will most likely represent abundant mRNAs (cf. Figure 4f).

Differences between the rates of re-annealing of abundant and rare cDNAs have been exploited to produce ‘equized’ cDNA libraries in which all species are equally abundant [28,29]. Takahashi and Ko [29] produced a mouse embryo equalized cDNA library by annealing PCR-amplified cDNA fragments, removing double-stranded DNA by hydroxyapatite chromatography, and PCR-amplifying the residual single-stranded cDNA. Two to three cycles of hybridization were sufficient to equalize the abundances of 26 cDNA species which varied in initial abundance from 2.4 to 0.00065%. We can model their data by modifying equation (2) so that $C_{\text{new}} = C_0 - B$, and using equation (10). Our model predicts that using as input variables the conditions from Takahashi and Ko [29], all species will be in equilibrium after two cycles of hybridization. These predictions are consistent with the experimental data (which suggested near but not complete equilibrium after two cycles), and give us further confidence that the real subtractive hybridizations can be modeled using a simple kinetic model.

**DISCUSSION**

In this report we have derived equations which allow us to model the purification of target sequences using the SHAP and SPAT techniques (with and without genomic sampling) and we have attempted to confirm the validity of our model by comparing its predictions with data from published genomic and cDNA subtraction experiments. We would have liked to have been able to make comparisons with more experimental observations, particularly in the case of cDNA subtractions. In nearly all the published cases, only a small proportion of the up-regulated cDNAs have been identified and the mRNA abundances determined [e.g. 2–4,6,27]. Moreover, in many of these reports, details such as relative abundance of up-regulated cDNAs within the subtractive library, and the mRNA levels in the (+) and (−) populations, were omitted. Most of the available data on cDNA subtractions were therefore unsuitable for testing the model because many of the input variables were not reported. Even in reports of genomic subtractions the data were often not complete enough to allow simulations to be run.

In our formulations we have made a number of assumptions, and their reliability depends on the extent to which these assumptions are justified. Britten and Davidson [23] have shown that during the renaturation of sheared, denatured double-stranded DNA, the observed rate of formation of DNA duplexes, as measured by hydroxyapatite chromatography, closely follows the expected second-order kinetics, at least until duplex formation is nearly complete. We have assumed that the kinetics of formation of driver–driver, driver–tester and tester–tester duplexes are the same [see equations (3) and (7)]. Biotinylination is unlikely to significantly affect the rate limiting step, since only ~1 nucleotide in 500 is modified [5]. However, disparities in the lengths of driver and tester strands plus additional, difficult-to-predict factors will result in deviations from ideal kinetics [11,20]. We have also assumed that all sequences are amplified to an equal extent by PCR. In fact, long sequences (>2 kbp) will be amplified less efficiently than shorter ones. Therefore, in cDNA subtractions, longer mRNA sequences are likely to be under-represented in the subtractive library. In genomic subtractions, where the DNA can be sheared or digested to short, relatively uniform fragments, the effects of differential PCR amplification caused by length disparities will be less, although it should be noted that such length differences form the basis for genomic sampling [9]. Variable PCR amplification efficiencies, caused by factors such as fortuitous primer binding to internal sequences, are even more difficult to predict, although the effects of any such variabilities may be minimal [29,30]. Despite these uncertainties, and the limited amount of data available for testing the model, the good general consistency between the predictions and the experimental data give us some confidence in the model.

For genomic subtractions, the model makes a number of predictions which are likely to have important practical implications in the design of experimental procedures. For organisms with small genomes (1C ≤ 5 × 10^8 bp), four to six cycles of
genomic subtraction using either SHAP or SPAT (without sampling) should prove equally efficient for cloning relatively small deletions (<5 kbp). Indeed for *Arabidopsis* or *Drosophila*, the target size is likely to be limited only by the requirement that the deletion be larger than the size of the DNA fragments. As genome size increases beyond 10^9 bp, the kinetics of hybridization start to become an increasingly important factor limiting enrichment of the target. This results in an advantage for SPAT compared to SHAP in the recovery of target sequences as C^+_k becomes the dominant term in equation (1). Predicted recoveries do not increase linearly with increasing target size, becoming smaller with increasing genome size for a target comprising a fixed proportion of the genome. For genomes with 1 C = 5 \times 10^8 bp, recovery of even large targets (>100 kbp) is limited by the kinetics of hybridization.

The model predicts that for the human or tobacco genomes, under conditions which maximize k, SHAP should require 9–10 cycles of hybridization to enrich a target of 5 kbp to near homogeneity, whereas using SPAT under similar conditions, only five–six rounds of hybridization are necessary (cf. Figures 3c and 3d). However, from an experimental perspective, this does not necessarily represent an advantage for SPAT. SHAP requires a PCR amplification step after each alternate hybridization cycle, whereas after each cycle, the number of PCR amplification steps will be similar under either procedure. Technically, the SHAP procedure is comparatively simple, requiring only an initial link ligation step followed by a series of hybridizations, phenol extractions and PCR amplifications [4]. In contrast, the SPAT technique [9] requires fresh linkers to be ligated to the residual DNA before each PCR amplification. Whether the kinetic advantage of SPAT for genomes of 1 C > 5 \times 10^8 outweighs the greater technical complexity is therefore uncertain. Genomic sampling [9] is potentially applicable to both SHAP and SPAT, and offers further kinetic enhancement. It will, to a first approximation, be equivalent to reducing the effective genome size to that of the sampled fraction. However, to ensure recovery of a specific target, the genome may have to be sampled several times using different restriction endonucleases [9]. Our model indicates that under favourable conditions, genomic sampling may be unnecessary even with large genomes, whereas Lisitsyn *et al.* [9] were unable to isolate targets in its absence. Possibly in real experiments, the cumulative effects, multiplied over several rounds of hybridization, of even small deviations from predicted recoveries may tip the balance against recovery.

Large, naturally occurring deletions (\geq 10^6 bp) have been found to be associated with genetic disorders [26], and deletions can be efficiently introduced experimentally using X-ray irradiation [24]. For mammalian cells, at the appropriate X-ray dose rate, 60–100% of mutations at selected loci resulted from deletions of between 1 and 14 kbp [25], the ideal size for genomic subtraction. Chromosome walking is the currently used routine technique for identifying and cloning both naturally occurring and experimentally induced mutations that confer interesting phenotypes. Compared to chromosome walking, genomic subtraction techniques offer great potential advantages, as they involve less time and effort. Here we provide a basis for the choice and optimization of protocols for cloning genomic deletions. By careful choice of the appropriate protocol it should be possible to routinely identify deletions from the human and other comparatively complex genomes.

For cDNA subtractions, the predictions made by the model are consistent with the experimental data. The model predicts that multiple rounds of hybridization will effectively reduce the levels of common background cDNAs to a very low level and enrich up-regulated sequences that are present at low initial levels. The extent to which cDNA abundances within a subtractive library can represent a complex population of mRNAs which vary in initial abundance and degree of up-regulation is debatable. Our model indicates how these parameters can affect the final representation. Where several mRNAs are up-regulated, the model predicts a complex relationship between the abundance of each species in the (+) and (−) populations and their relative abundance after several rounds of subtraction. Thus, a clone might be abundant in a subtractive cDNA library because it was derived from either a highly abundant mRNA with a moderate degree of up-regulation, or from a very low abundance sequence that was highly up-regulated. In practice, the model indicates that subtractive cDNA cloning will successfully isolate low abundance sequences provided their degree of up-regulation is greater than that of other up-regulated sequences. This point is illustrated by the levels, predicted and observed [4], of *edeA* and *edeB* sequences in the subtractive library from *N. glauca*.

Our analysis of cDNA subtractive hybridization has used the tobacco leaf cell as a model. However, the mRNA populations of a variety of eukaryotic cells show rather similar mRNA abundance classes [13]. The outcomes of subtractive hybridization should therefore not be expected to differ much between organisms, and six to seven rounds of hybridization should prove adequate in all cases. Since the number of rounds of hybridization required depends on the number of copies per cell of the least abundant common mRNAs, this prediction of the model is likely to hold true even in tissues with unusual distributions of mRNA species.

ACKNOWLEDGEMENTS

We should like to thank Dr J. A. Jackson for helpful suggestions and ideas. Thus work was supported by a grant from the Glasgow University New Initiatives Fund.

REFERENCES