The Mathematical Model of Subtractive Hybridization and its Practical Application.

Olga D. Ermolaeva, Sergey A. Lukyanov and Eugene D. Sverdlov.

Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry, RAS, 16/10 Miklukho-Maklaya, Moscow, 117871, Russia.
Phone & Fax: 7-095-330-6538. E-mail: ermolaevaola@glas.apc.org

Abstract

A novel theory of subtractive hybridization including (or based on) the kinetic model of this process was proposed. A computer program modeling the process of subtraction was developed. Basing on the theory, a novel method of subtractive hybridization was proposed allowing routine comparison of genomes and products of genome expression. The method was applied to studies of the genetic mechanisms of embryogenesis, regeneration, cell differentiation and tumor transformation.

Introduction

Isolation of DNA fragments responsible for the phenotypic differences is one of the most important problems of molecular biology. Method of subtractive hybridization plays a prominent role among the methods of isolation of differences of closely related genomes or products of genome expression (Sverdlov 1993). This method allows one to construct a library of differences of the genomes of compared organisms directly, resulting in fast obtaining of structural information of the genome differences for the purposes of subsequent functional analysis.

This analysis may lead to understanding of genetic causes of the differences between species, racial and individual differences, molecular causes of hereditary deceases, tumor cell transformation and metastasis.

There are a number of examples of successful application of the subtractive hybridization.

This method was used in a work that played a crucial role in localization of the Duchenne muscular dystrophy gene (Kunkel et al. 1985) and for the construction of the library enriched with Y-chromosome specific sequences (Lamar & Palmer 1984). A number of genes that change their level of expression in different tissues or at different stages of embryogenesis were identified. For example, subtractive hybridization was used for identification of genes coding for the receptor proteins that are present on the surface of T-lymphocytes but absent from the surface of B-lymphocytes (Hedrick et al. 1984) and also in a search for factors controlling the process of ageing (Hara et al. 1993).

Nevertheless, this method is not yet developed well enough for solving such problems as search for small differences between complex genomes and isolation of the rare transcripts in cDNA subtraction.

The analysis of literature shows that the kinetic characteristics of the process of hybridization that forms the basis for the subtractive hybridization techniques are not usually taken into consideration. The choice of strategy of subtraction and experimental conditions is usually based on intuition and not on a theoretical analysis. The application of the relevant theory in these cases should lead to much better experimental results.

Thus, the goal of this work was to develop a theory of subtractive hybridization. Such a theory should allow the subtractive hybridization to be turned into a routine process of obtaining structural information about differences of genomes or products of genome expression. This work was divided into 3 stages:

Development of the kinetic model of subtractive hybridization; which should serve as a basis for the theoretical analysis of the method of subtractive hybridization;

Development of software tools for the practical utilization of the above theory;

Analysis of different applications of subtractive hybridization basing on the developed theory.

The Formulation of the Problem

The most traditional application of subtractive hybridization is the isolation of sequences that are present in one of DNA populations but absent in another, for example, as a result of genome mutation (deletion). Such sequences are called **"target"** sequences. DNA that contain target sequences will be called **"tracer"** DNA, DNA that is identical to tracer with exception of target sequences will be called **"driver"** DNA.

DNA is fragmented (for example by restriction endonuclease digestion) and tracer DNA is mixed with an excess amount of driver DNA. The resulting mixture is denatured, then the DNA is allowed to reassociate. Tracer fragments that are present in driver as well will reassociate mostly with their driver counterparts, since they are present in great excess. Target sequences will reassociate

with each other. Duplexes that contain DNA fragments common for tracer and driver are removed. The remaining DNA is enriched with target molecules. This DNA is cloned and analyzed. This type of the target will be called the **"absolute"** target.

The practical implementation of such an approach requires a way to estimate the achieved enrichment and content of target sequences in the resulting library.

Subtractive hybridization may be also used for isolation of sequences that are present both in driver and in tracer but in different abundance. Such a problem often arises during subtraction of cDNA libraries for the determination of the differences in levels of gene expression in closely related cell types. This type of the target will be called the **"copied"** target.

Subtractive hybridization may be used also for the isolation of homologous sequences present both in driver and tracer. This type of target will be called **"homologous"** target.

All reported methods for the genome subtraction utilized the same strategy - the subtraction of double-stranded tracer and double-stranded driver. In principle, other strategies may be employed. For example, it is possible to use single-stranded tracer and single-stranded driver that is complementary to tracer. In this case target sequences do not have counterparts neither in tracer nor in driver and their concentration remains constant. So it is possible to reach greater degree of enrichment after the removal of duplexes.

In this work we have analyzed the process of subtractive hybridization for the following cases:

absolute target;

copied target;

homologous target.

For all target types we have analyzed 5 subtraction strategies:

1. double-stranded driver and double-stranded tracer;
2. single-stranded tracer and double-stranded driver;
3. single-stranded tracer and complementary single-stranded driver;
4. double-stranded tracer and driver that is unable for reassociation;
5. double-stranded tracer and single-stranded driver.

The Kinetic Model of Subtractive Hybridization

The proposed kinetic model of subtractive hybridization is based on the assumption that the rate of the formation of driver-driver, tracer-tracer and driver tracer duplexes is described by the second order equation (Sverdlov & Ermolaeva 1993, Sverdlov & Ermolaeva 1994). It consists of a system of differential equations describing the kinetics of DNA reassociation in solution. A number of factors (such as salt concentration, temperature, DNA fragment length etc.) are also taken into consideration.

Let us introduce the following designations:

D, T, U, W - molar concentrations of fragments of driver, fragments of tracer that are present in driver, fragments of target in tracer, and fragments of target in driver, respectively;

d, s - indices, corresponding to double-stranded and single-stranded DNA, for example U^s, T^d, D^s;

$U^s(t)$, $D^s(t)$ - the concentrations as a function of time;

U_0, T_0 - the initial concentrations of single-stranded DNA;

E_s, E_d - the enrichment in single-stranded and double-stranded DNA fractions respectively;

R, $M^{-1} \cdot s^{-1}$ - the rate of reassociation.

$R \approx 10^6$ $M^{-1} \cdot s^{-1}$ for the fragments with length $\approx$ 500 nucleotides in 0.18 M NaCl at optimal temperature equal to T_m- 25°C (Britten & Davidson 1985). R_1 - the reassociation rate for the fragments that do not share 100% homology.

The kinetics of the process of subtractive hybridization is described by the following systems of differential equations:

(1):

$$\begin{cases} \dfrac{dD^s(t)}{dt} = -a_{00}RD^s(t)D^s(t) - a_{01}RD^s(t)T^s(t) \\ \dfrac{dD^d(t)}{dt} = a_{10}RD^s(t)D^s(t) \\ \dfrac{dT^s(t)}{dt} = -a_{20}RT^s(t)T^s(t) - a_{21}RD^s(t)T^s(t) \\ \dfrac{dT^d(t)}{dt} = a_{30}RT^s(t)T^s(t) \end{cases}$$

$$D^s(0) = D_0,\ D^d(0) = 0,$$
$$T^s(0) = T_0,\ T^d(0) = 0.$$

CLASS	A number of individual mRNA sequences	A number of copies of individual mRNA sequences per cell.	% content of a given mRNA sequence relative to total mRNA.
High-abundant	12	7200	2.4
Medium-abundant	330	250	$8.24 \cdot 10^{-2}$
Low-abundant	6400	21	$6.92 \cdot 10^{-3}$

Table 1. Different classes of cytoplasmic poly(A)$^+$-mRNA in mammalian cells (Galau et al. 1977).

(2):

$$\begin{cases} \frac{dW^s(t)}{dt} = -b_{00}RW^s(t)W^s(t) - b_{01}RW^s(t)U^s(t) \\ \frac{dW^d(t)}{dt} = b_{10}RW^s(t)W^s(t) \\ \frac{dU^s(t)}{dt} = -b_{20}RU^s(t)U^s(t) - b_{21}RW^s(t)U^s(t) \\ \frac{dU^d(t)}{dt} = b_{30}RU^s(t)U^s(t) \end{cases}$$

$$W^s(0) = W_0,\ W^d(0) = 0,$$

$$U^s(0) = U_0,\ U^d(0) = 0.$$

$$E_s(t) = \frac{U^s(t)}{T^s(t)};$$

$$E_d(t) = \frac{U^d(t)}{T^d(t)}.$$

For each case the coefficients a_{ij}, b_{ij} of the systems of differential equations (1), (2) are different and equal to 0 or 1, since, for example, if driver and tracer are single-stranded, driver is able to reassociate with tracer but is not able to reassociate with itself. In case of absolute and copied target R_1=R and in case of absolute target W_0=0.

System (1) has solutions in elementary functions assuming that the concentration of driver is much greater than the concentration of tracer. In some cases system (2) has no solutions even if the concentration of driver greatly exceeds that of tracer. This made us to write a SUBTRACT software package that implements the modeling of subtractive hybridization. The user-friendly interface of this program developed by Mark Wagner (Ermolaeva & Wagner 1995) allows the casual user to choose optimal experimental conditions and strategy of subtraction. The program was developed on Sun Microsystems hardware using the standard C language. The Motif widget system (version 1.0) was used to implement the independent windows and controls. The actual drawing system was mplemented using the X11 Window system (Release 5). This gives SUBTRACT the ability to be run over the network, with the code being run on one workstation and the interface being displayed on any system that can display X Windows.

Analysis of Different Subtraction Approaches

The SUBTRACT program allowed us to perform additional analysis of subtraction of genomes and cDNA libraries leading to unexpected conclusions and to propose more effective approaches to subtractive hybridization.

a) The Subtractive Hybridization of Transcripts

The subtractive hybridization approach is widely used for isolation of transcripts that are present in some cells and absent in other, closely related, cells. Nevertheless this method is not yet well suited for the isolation of transcripts that are present in very small amounts (Lukyanov et al. 1994). One of the major problems in isolation of the rare transcripts is the fact that different mRNA classes are present in cells in different amounts.

The cellular mRNAs can be subdivided into 3 classes (see table 1): high-abundant, medium-abundant and low-abundant (Galau et al. 1977). Molecules of the first class are produced by 10 to 20 genes, mammalian cells contain a few thousands copies of each of these transcripts. Molecules of the second class are transcribed from several hundred genes and cells typically contain several hundred copies of each of these transcripts. Most genes, however, produce only a few transcripts per cell.

The cDNA subtraction strategies proposed earlier were not developed specifically for search for a targets of the specific mRNA class. It is obvious that search for members of different classes calls for different strategies. Thus it would be desirable to develop specific strategy optimized for a given task, for example, for the search for rare transcripts.

We have analyzed the most widely used approach to cDNA subtraction (Sverdlov & Ermolaeva 1994). This

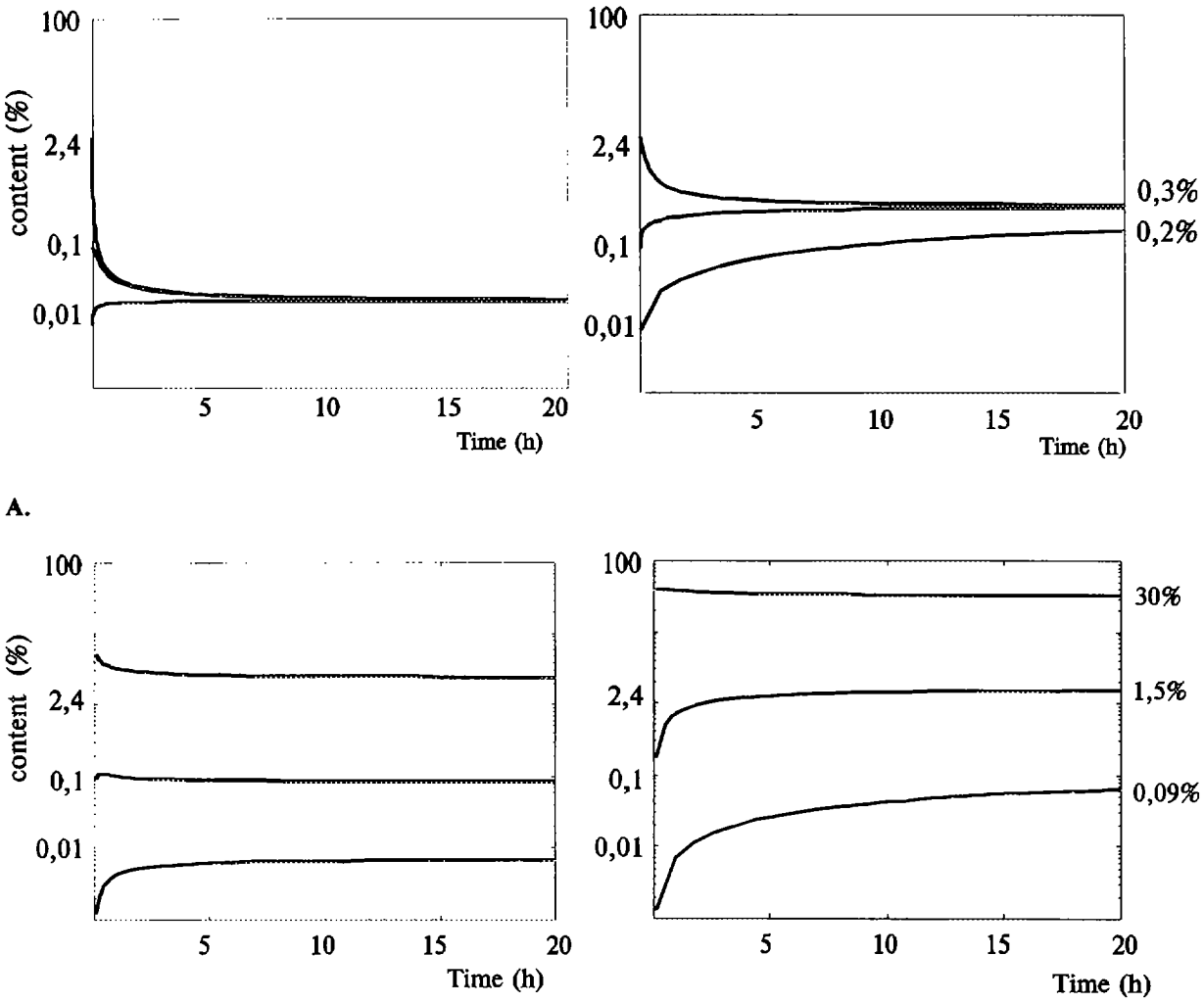

Figure 1. The content of target sequences in single-straned (**A**) and double-stranded (**B**) DNA fractions after the first round of subtraction. **Time** - the hybridization time (in hours). **Content** - the content of the particular sequences belonging to high-, medium- and low-abundant mRNA classes (top to bottom, relative to total DNA amount).

approach is used for the search for mRNA molecules that are present in one type of cells but absent from another type and is based on subtraction of double-stranded tracer against double-stranded driver.

Our calculations show that the resulting enrichment will be different for different classes of target molecules (whether target molecules belong to high-abundant, medium-abundant or low-abundant class).

Fig. 1A presents the results of calculations for single-stranded fraction of tracer showing normalization of concentrations of different types of molecules for this fraction. After 20 hours of hybridization content of highly-abundant target is 8 times less than in the beginning, of middle-abundant target - 3 times more and of low-abundant target - 20 times more.

Fig. 1B presents similar results for the double-stranded fraction of the tracer. In this fraction we can see a dramatic increase of the content of high abundant target, which constitutes up to 30% of total DNA. The concentration of low-abundant target is 9 times more than in the beginning, but still is quite low.

The obvious conclusion is that the double- and single stranded fractions of tracer should be separated after the first round of subtraction. The double-stranded fraction can be cloned producing a library with high content of targets originating from high-abundant mRNA. The single-stranded fraction should be used in subsequent subtraction rounds for isolation of targets originating from low-abundant mRNA.

Fig. 2 presents the content of target sequences in double- and single stranded fractions of tracer after second

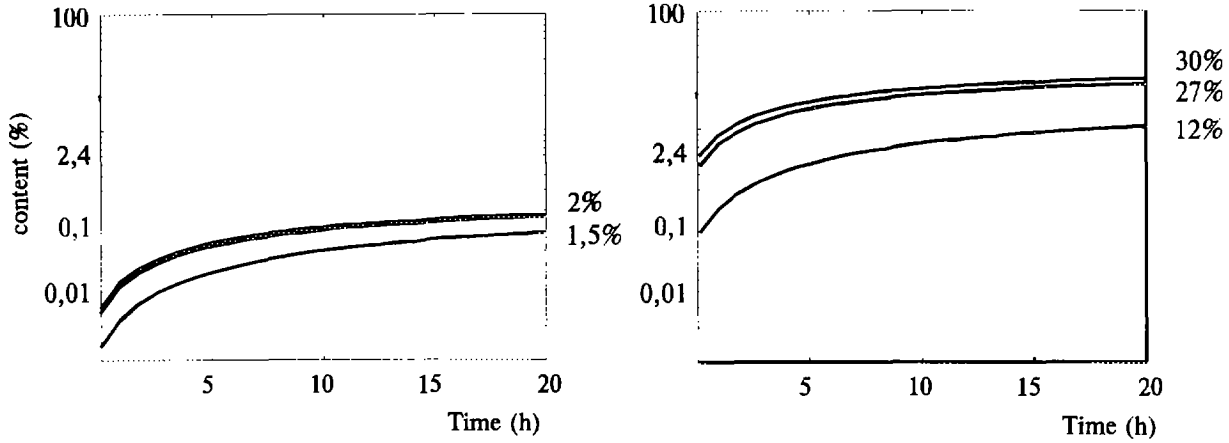

Figure 2. The content of target sequences in single-stranded (left) and double-stranded (right) fractions of DNA after second round of subtraction. **Time** - the hybridization time (in hours). **Content** - the content of the particular sequences belonging to high-, medium- and low-abundant mRNA classes (top to bottom, relative to total DNA amount).

round of subtraction. In this case double-stranded fraction is highly enriched in target sequences originating from all 3 mRNA classes.

About 30% of all DNA fragments in this fraction originate from high-abundant mRNA, 27% - from medium-abundant and 12% - from low-abundant.

Thus, 2 cycles of subtraction should suffice for obtaining a cDNA library with high enough representation of rare target sequences. In order to achieve this goal one should save the single-stranded fraction of not reassociated

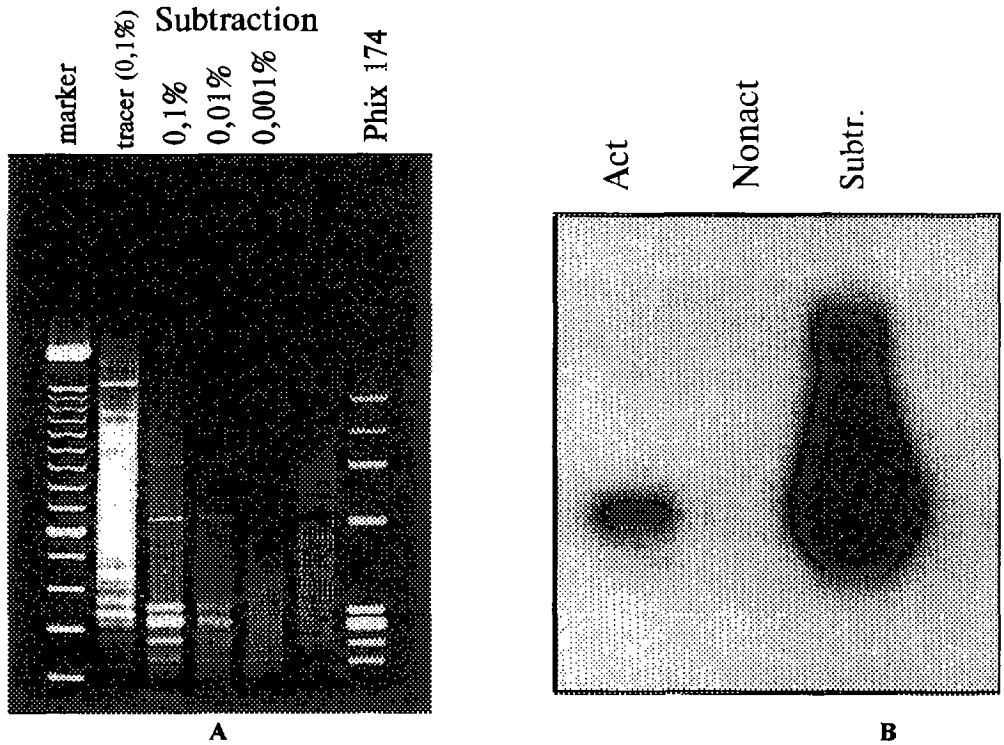

Figure 3. **A.** Lane 1 - DNA size marker (100 bp ladder) The next lane - tracer DNA (muscle cDNA containing 0.1% of ϕX174 DNA digested with HaeIII) followed by three lanes with products of subtraction of muscle cDNA containing phage DNA in amount equivalent to 100, 10 and 1 copies per cell. The last lane - HaeIII digest of ϕX174 DNA. **B.** Lane 1 - tracer (cDNA from Jurkat cells activated by PGA), lane 2 - driver (cDNA from not-activated cells), lane 3 - subtraction results.

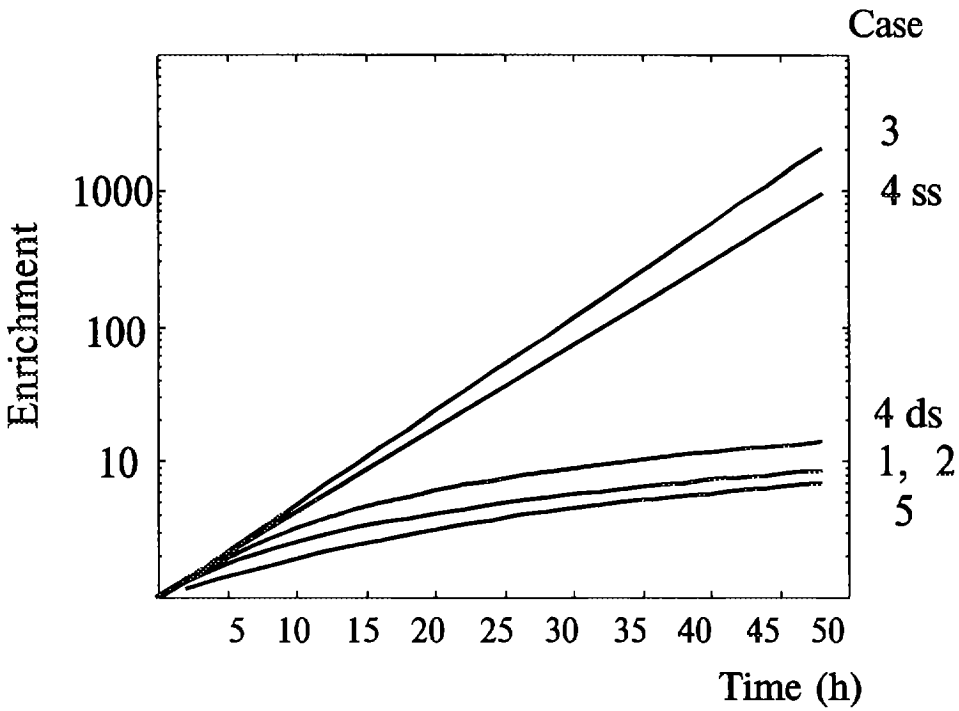

Figure 4. The expected enrichment of target for different strategies of subtractive hybridization. **Time** - the hybridization time (in hours). **Enrichment** - the enrichment of the target sequence. **1** - double-stranded driver and double-stranded tracer; **2** -single-stranded tracer and double-stranded driver; **3** - single-stranded tracer and complementary single-stranded driver; **4** - double-stranded tracer and driver that is unable for reassociation; **5** - double-stranded tracer and single-stranded driver.

tracer molecules after the first round in subtraction for use in additional subtraction rounds. In following rounds a double-stranded fraction of tracer should be saved.

Isolation of single-stranded molecules can be achieved by different means, for example by physical separation of single- and double stranded DNA after hybridization. However a specific amplification of single-stranded DNA and not double-stranded DNA seems to be more effective.

One of variants of the latter approach was used in a strategy for the isolation of rare transcripts that we developed on the base of proposed theory (Lukyanov et al. 1994). The PCR primers of particular structure that allow us to amplify only single-stranded fraction of tracer were used. We performed a theoretical analysis of the proposed strategy and estimated the expected enrichment.

The experimental results of this scheme are presented in Fig. 3. Phage DNA was added to muscle cDNA as a target (Fig. 3a) in an amount corresponding to 1, 10 and 100 copies per cell and can be considered to belong to low- and medium-abundant mRNA classes. The resulting enrichment was about 100 times for the first case and at least 1000 times for the second case, in agreement with the theoretical prediction.

The practical applicability of this scheme was additionally verified by isolation of gene transcripts that are synthesized in Jurkat cells upon PGA activation, such as IL-2 receptor. The results of the comparison of content of IL-2 receptor cDNA prepared from activated and normal cells and subtraction products are presented in Fig. 3b.

b) Genome Subtraction

We also applied the proposed kinetic model to theoretical analysis of the process of genome subtraction (Sverdlov & Ermolaeva 1993). Despite the fact that the subtractive hybridization is widely used for the comparison of genomes, no reliable methods for the isolation of small differences between complex genomes were developed.

The most widely used strategy for the genome subtraction is the hybridization of double-stranded driver and double-stranded tracer. The comparison of the enrichment for all strategies of subtraction is presented in Fig. 4. Our data show that the use of single-stranded tracer and complementary single-stranded driver should results in higher enrichment than the use of alternative strategies. In this case after 30 hours of hybridization the enrichment of target should be about 120 times, whereas the traditional strategy should result only in 6-fold enrichment.

We propose an optimal strategy for the genome subtraction which includes the use of single-stranded tracer and single-stranded driver complementary to tracer. The development of this strategy should allow for the direct isolation of small differences between complex genomes without the preliminary stages which bring the danger of losing the difference to be sought.

Conclusion

Our experience in application of the proposed theory to several cases of subtractive hybridization shows that it adequately describes the process of subtraction experiment. One of the major advantages of this theory is it's universality. It can be applied to determination of the experimental conditions for existing strategies of subtractive hybridization and for the analysis of the viability of the novel strategies for the search of specific differences as well.

Acknowledgments

The authors are grateful to K. Lukyanov and N. Gurskaya for kind permission to use figure 3. We also would like to thank Dr. S. Volik for translation of manuscript.

References

Britten, R. J.; and Davidson, E. H. 1985. Hybridisation strategy. In Hames, B. D.; and Higgins S. J. (eds), *Nucleic acid hybridisation*. IRL Press, Oxford-Washington DC, 3-14.

Ermolaeva, O. D.; and Wagner, M. C. 1995. SUBTRACT, a computer program for modeling the process of subtractive hybridization. *CABIOS* 11: 457-462.

Galau, G. A.; Klein, W. H.; Britten, R. J.; and Davidson, E. H. 1977. Significance of rare mRNA sequences in liver. *Archives of Biochemistry and Biophysics* 179: 584-599.

Hara, E.; Yamaguchi, T.; Tahara, H.; Tsuyama, N.; Tsurui, H.; Ide, T.; and Oda K. 1993. DNA-DNA subtractive cDNA cloning using oligo$(dT)_{30}$ -Latex and PCR: identification of cellular genes which are overexpressed in senescent human diploid fibroblasts. *Anal. Biochem.* 214: 58-64.

Hedrick, S. M.; Cohen, D. I.; Nielsen, E. A.; and Davis, M. M. 1984. Isolation of cDNA clones encoding T cell-specific membrane-associated proteins. *Nature* 308:149-153.

Kunkel, L. M.; Monaco, A. P.; Middlesworth, W.; Ochs Hans, D.; and Latt, S. A. 1985. Specific cloning of DNA fragments absent from the DNA of a male patient with an X chromosome deletion. *Proc. Nat. Acad. Sci. USA* 82: 4778-4782.

Lamar, E.; and Palmer, E. 1984. Y-encoded, species-specific DNA in mice: evidence that the y chromosome exists in two polymorphic forms in inbred strains. *Cell* 37: 171-177.

Lukyanov, S. A.; Gurskaya, N. G.; Lukyanov, K. A.; Tarabykin, V. S.; and Sverdlov, E. D. 1994. Highly efficient subtractive hybridization of cDNA. *Bioorg. Khim.* 20:701-704.

Sverdlov, E. D. 1993. Subtractive hybridization - technique of extraction of DNA sequences discriminating between two closely related genomes. *Mol. Gen. Microbiol. Virusol.* 1-12.

Sverdlov, E. D.; and Ermolaeva, O. D. 1993. Subtractive hybridization. Theoretical analysis and a principle of the "trapper". *Bioorg. Khim.* 19: 1081-1088.

Sverdlov, E. D.; and Ermolaeva O. D. 1994. Kinetic analysis of subtractive hybridization of transcripts. *Bioorg. Khim.* 20: 506-514.