Identification of cDNA Sequences by Specific Oligonucleotide Sets. Computer Tool and Application

Kolchanov N.A., Vishnevsky O.V., Babenko V.N., Kel A.E., Shindyalov I.N.#)

Institute of Cytology and Genetics, Siberian Branch, Russian Academy of Sciences, 630090,Novosibirsk, Russia, e.mail: kol@cgi.nsk.su, fax: (3832) 356558

#) Columbia University, 630 W. 168th Street, New York, NY 10032 USA, email: shindyal@cuhhca.hhmi.columbia.EDU

Abstract

A computer tool has been developed for revealing sets of oligonucleotides invariant for isofunctional families of DNA (RNA) and for using these in functional identification of nucleotide sequences. The tool allows one to: build up vocabularies of invariant oligonucleotides for the families of isofunctional nucleotide sequences; assess significance of the vocabularies; identify nucleotide sequences with the vocabularies of invariant oligonucleotides; determine the most effective identification parameters to minimize first and second type errors; assess the efficiency of identification of individual isofunctional families with the oligonucleotide vocabularies; determine the evolutionary characteristics of the families of isofunctional sequences on which vocabulary volume depends. Based on the system mentioned, we have analyzed a total of 322 protein-encoding gene families and have built up sets of invariant oligonucleotides, or again, oligonucleotide vocabularies that are characteristic of gene families and subfamilies. Identification of nucleotide sequences belonging to these families with the sets of invariant oligonucleotides revealed has been shown. Under the most effective identification parameters, the first type error (false negative) on control (independent) data was 10-15%, the second type error (false positive) was just 1-2 redundant sequences per sequence being examined. As has been shown, the volume of a vocabulary of invariant oligonucleotides depends on the percentage of variable positions in the multiple alignment within a family.

Introduction

Investigation of the eucaryotic genomes that are billions of base pairs in length has encouraged the development of new approaches for studying the structural organization of genomic DNA, which are based on various oligonucleotide techniques. These are: the methods of sequencing by hybridization to sets of short oligonucleotides (Khrapko et al.,1989; Kuznetsova et al., 1994; Strezoska et al., 1991), hybridization with oligonucleotides for detection of polymorphic sites (Davies, 1986; Saiki et al., 1986), in particular, single base substitutions (Khrapko et al., 1991), and oligonucleotide probes for screening cDNA libraries (Suggs et al., 1981) and for ordering clone libraries (Hoheisel, 1994). Recently, the oligonucleotide composition of sequences has been used for phylogenetic inferences (Solovyev et al., 1993). These works illustrate the opportunities of the oligonucleotide analysis.

The oligonucleotide techniques can be successfully applied in classification of cloned DNA (genomic fragments of DNA or cDNA) into functional classes by hybridization with specific sets of short oligonucleotides. Functionally significant classes of sequences include the genes coding for proteins and various types of RNA, gene regions coding for the most common protein motifs (zinc finger, homeobox etc.), functional regions providing the structural organization of chromatin, including the sites of

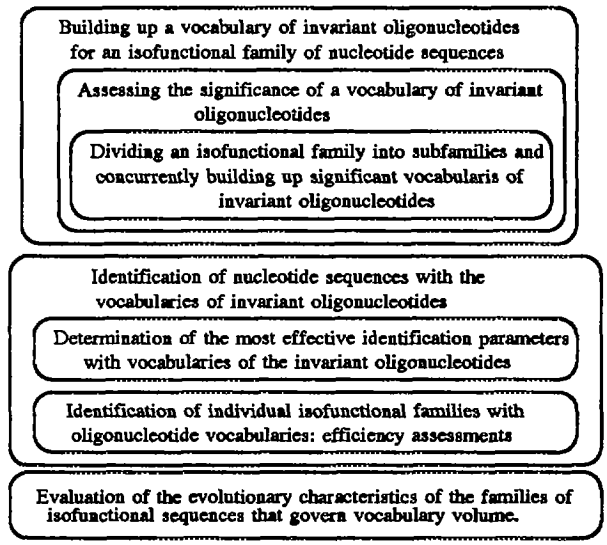

Figure 1. A computer tool for classification and identification of nucleotide sequences with vocabularies of invariant oligonucleotides.

association with the matrix, telomeric and centromeric repeated sequences; promoters and other signal sequences that control gene expression at transcription, processing, splicing, translation.

Techniques that allow fragments of cloned DNA to be classified by functional region type on the basis of hybridization with a limited number of specific oligonucleotides, i.e. without total sequencing by any of the traditional methods, can considerably reduce the amount of time spent on the experimental study of eucaryotic genomes, in the first turn - on functional mapping. One of the important points on the way towards such techniques is the development of theoretical approaches and software that allow the investigator to build up sets of specific oligonucleotides for effective decision as to what type of functional region a nucleotide sequence should belong to. Also it is required that any resulting set of oligonucleotides provides the least possible errors at functional classification of the respective functional region.

Here we report a computer tool that we have developed to build up sets of oligonucleotides invariant for isofunctional cDNA families (Figure 1). We have analyzed with it 322 families of protein-encoding sequences and built up sets of invariant oligonucleotides, or oligonucleotide vocabularies, that are characteristic of these families. How the nucleotide sequences of these families can be identified with the use of such invariant oligonucleotide sets has been shown, too. Under the most effective values of the identification parameters, the first type error (false negative) on control (independent) data was 10-15%, and the second type error (false positive) was just 1-2 redundant sequences per sequence under trial.

Materials and methods

Sequences used for analysis.

Original data were taken from the database on 322 families of isofunctional genes coding for proteins (Shindyalov et al., 1993). Each family consists of coding sequences with not less than 50% homology. The average number of sequences in the family was M=12, although some families comprised from 6 (6-phosphogluconate dehydrogenase) to 103 sequences (nucleoprotein). The average sequence length for a family was N=730 bp, the least being 84 bp (kappa chain V-region immunoglobulin) and the most being 3126 bp (ATPase subunits).

Vocabulary of invariant oligonucleotides for a family of isofunctional nucleotide sequences.

Let $S = \{ S^j \}$ be a family consisting of M nucleotide sequences S^j of length L each ($j=1,..., M$). Let W^j be the complete set of the oligonucleotides of length l each, contained in the j-th sequence of that family. A vocabulary of the invariant oligonucleotides of the family S is defined as the intersection of the oligonucleotide vocabularies of all sequences of the family and is hereinafter called "the oligonucleotide vocabulary of the family S", or just "vocabulary" for short:

$$W = \bigcap_{j=1}^{M} W^j$$

Let $R > 0$ be the number of oligonucleotides in the vocabulary W. Note that for combinatorial reasons, even a group of random nucleotide sequences may have a set of common oligonucleotides. Therefore, it is of importance to assess significance for the volume of the vocabulary W containing R oligonucleotides of length l each that occur in all the sequences of the family S, each of which is of length l.

Assessing significance for the oligonucleotide vocabulary

To have the simplest case of such assessment, consider a pair of random sequences of lengths L and N with the nucleotide frequencies of P_1^A, P_1^T, P_1^G, P_1^C and P_2^A, P_2^T, P_2^G, P_2^C respectively.

Then the frequency of the same nucleotide observed in the two sequences is

$$p = P_1^A P_2^A + P_1^T P_2^T + P_1^G P_2^G + P_1^C P_2^C \qquad (1)$$

Here the frequency of the coincidence of a definite oligonucleotide of length l in the two sequences is:

$$p = p^l \qquad (2)$$

The probability of K oligonucleotides of length l being identical in the two sequences is:

$$P(K) = C_{LN}^{K} p^K (1-p)^{LN-K} \qquad (3)$$

The probability of more than K oligonucleotides being identical in the two sequences of the length L and M respectively is:

$$P(r \geq K) = \sum_{i=K}^{LN} C_{LN}^{i} p^i (1-p)^{LN-i} \qquad (4)$$

We find K_0 such that

$$P(r \geq K_0) \leq \alpha \text{ and } P(r \geq K_0 - 1) > \alpha \qquad (5)$$

So specified, Ko represents the upper limit of the interval of significance for the average expected number of the oligonucleotides of length l absolutely identical in two random sequences of length L and N (in the particular case, when the sequences are equal in length, $LN=L^2$).

Formulae (3) and (4) are valid in the case of independent (i.e. nonoverlapping) oligonucleotides. To estimate the expected probabilities (3), (4) in case of dependence, Monte Carlo simulation has been performed for the lengths of sequences $L=N=100, 200, \ldots 2000$ for $\alpha = 0.05$, 0.01, 0.001 and 0.0001. As turned out, for the indicated values of N, L and α, the value Ko, which takes account of overlapping, can be determined with a good accuracy from a linear correction of an 'independent' model (3)-(5):

$$Ko' = Ko + 3 \qquad (5a)$$

All the results to be presented below were obtained on the basis of the upper limit of the interval of significance corrected in line with (5a).

Note that the more sequences in the family S, the lower the probability of the vocabularies being identical by chance. That is why, *ceteris paribus*, the upper limit of the interval of significance Ko^* for three and more sequences satisfies the inequality:

$$Ko^* < Ko \qquad (6)$$

It means that Ko obtained from (2) - (5) is overestimation, if $M>2$. Thus, while assessing significant volume for the vocabulary W containing R oligonucleotides that are identical in M sequences of the family S, the following condition is checked:

$$R > Ko \qquad (7)$$

As Ko is overestimation, criterion (7) is still in force when we wish to build up an oligonucleotide vocabulary of a significant volume. Thus, if criterion (7) holds true, we may acknowledge that M sequences of the family S with the significance level α have an oligonucleotide vocabulary of the volume R.

Note that in pursuit of estimates (2)-(7) we did not use any information on homology between the sequences or on their alignment. As first stage of analysis, we built up a vocabulary W of invariant oligonucleotides for the sequences of the family S. Then we assessed its significant volume R by using (2)-(7). Whenever condition (7) was satisfied, the vocabulary W continued to participate in further analysis. Whenever it was impossible to achieve any significant vocabulary for the family S, we knew that the family was not homogeneous, i.e. that the family contained two or more subfamilies of significantly different sequences. It was therefore required to divide the original family into subfamilies, whose vocabularies were significant.

Dividing an isofunctional family into subfamilies and building up significant vocabularies of invariant oligonucleotides

Let W^j be the vocabulary for the j-th sequence. The number, d_{ij}, of oligonucleotides identical in the vocabularies is supposed to be a measure of similarity between the i-th and j-th sequences. Create the matrix $D_M = /d_{ij}/\ i,j = 1,\ldots,M$. Then take the following steps.

I. Find the greatest element of the matrix D_M that corresponds to the i^*-th and j^*-th sequences. This choice means that the two sequences have the most similar vocabularies:

$$d_{i^*j^*} = \max_{ij} d_{ij}, \qquad (i < j)$$

Table 1. Parameters of some families of isofunctional genes and the characteristics of their significant oligonucleotide vocabularies

Family name	Length (bp)	Homology (%)	Sequences number (qty)	Observed invariant oligonucleotides, R, (qty)	Expected invariant oligonucleotides (qty)	Upper limit of significance Ko (qty)	Percentage of invariant oligonucleotides, C (%)	Variable positions, Q (%)	Unevenness G
factor ix	634	64	5	68	11	19	11	26	0.89
alcaline phosphatase	1414	92	7	164	38	53	12	24	0.77
egg-laying hormone	631	77	5	548	11	19	87	1.6	0.72
flagellin	1711	87	6	322	53	71	19	29	1.2
glucocorticoid receptor	2218	86	6	725	84	107	33	16	1.12
phosphoglycerate kinase	1243	70	7	606	30	44	49	21	2.8
pilin	469	78	5	80	8	15	17	38	1.5
env-gene	2392	77	9	134	97	122	5.6	43	0.92
preapolipoprotein E	907	78	5	102	18	29	11	33	1
collagen	811	60	5	47	15	25	5.8	53	0.85
cell surface glicoprotein	706	79	10	30	13	22	4.2	36	0.88
ml protein	757	94	6	231	14	23	30.5	14	0.93

II. The following condition should be satisfied:

$$d_{i^*j^*} \geq Ko, \quad (8)$$

where Ko is the upper limit of significance as in (2)-(7). If satisfied, then the i^*-th and j^*-th sequences are significantly similar in that the i^*-th and j^*-th oligonucleotide vocabularies are similar.

III. Now build up a vocabulary of the oligonucleotides that are common to the i^*-th and j^*-th vocabularies: $W_{i^*j^*} = W_{i^*} \cap W_{j^*}$

Bring the i^*-th and j^*-th sequences into one current subfamily $U_{i^*j^*}$ with the vocabulary $W_{i^*j^*}$. In fact, any current subfamily as this is the union of the sequences with the most similar vocabularies.

IV. Transform the matrix D_M of dimensions MxM into the matrix D_{M-1} of dimensions $(M-1)x(M-1)$ by the following procedures:

a) Cross out the j^*-th column and the i^*-th row from the matrix DM to reduce its dimension by 1;

b) Replace the elements of the i^*-th column and the j^*-th row by the values of the distances between the current subfamily $U_{i^*j^*}$ and the other sequences (current subfamilies) of the family S;

c) The distance between the current subfamily Ui^*j^* and the r-th sequence is defined as the number of identical oligonucleotides in the vocabularies $W_{i^*j^*}$ and W_r.

The matrix D_{M-1} defines the distances between $M-1$ vocabularies. $W_{i^*j^*}$ is the vocabulary related to the current subfamily of two similar sequences and the others are related to individual sequences of the family S.

V) Repeat the procedures from (I) through (IV).

Every current iteration reduces the dimensions of the matrix and increases the number of current classes and (or) the number of the sequences in them with similar oligonucleotide vocabularies. The classification procedure terminates as soon as condition (8) cannot be satisfied for any i^*j^*. It implies that the original family S has been divided into classes (subfamilies), each of which consists of sequences with similar vocabularies, whereas the sequences related to different classes have different vocabularies. In fact, this algorithm represents an implementation of the UPGMA technique of grouping with a particular way of calculating the distance between subfamilies and a specified criterion for termination of the grouping procedure.

Results and discussion

Building up the significant vocabularies of invariant oligonucleotides for the isofunctional families of protein-encoding genes.

For the purposes of our analysis, we took 8 bp long oligonucleotides (l=8) as the most frequently used in various oligonucleotide techniques. The calculations were made with a level of significance of 0.01. Of 322 families analyzed, 172 were homogenous, i.e. we have succeeded in building up significant vocabularies of invariant oligonucleotides for them without referring to division into subfamilies (Table 1).

For example, the factor IX gene family consists of 5 sequences 634 bp in length on the average at average homology of 64%. The vocabulary W contains R=68 invariant oligonucleotides. The upper limit Ko=19 at a significance level of 0.99%. Thus, in line with (7), a significant vocabulary has been built up for the family in question. For the other families presented in Table 1, the number of observed invariant oligonucleotides exceeded Ko.

For 150 families out of the 322, it was not till we had divided them into subfamilies that we had succeeded in building up significant vocabularies (see examples in Table 2). For instance, for the a-actin gene family which consists of 12 sequences, the registered number of invariant oligonucleotides R=27 which is less than Ko=38. Therefore, no significant vocabulary might be successfully built up for the family as a whole. By applying the division procedure, we got two subfamilies each consisting of 6 sequences. At Ko=38, the vocabulary W for the former subfamily has R=47 invariant oligonucleotides, whereas the vocabulary W for the latter has R=114. Thus, the respective significant vocabularies

Table 2. Division of some families of isofunctional genes into subfamilies for setting up nonrandom oligonucleotide vocabularies.

Original family name	Sequences number (qty)	Invariant oligonucleotides R,(qty)	Upper limit of significance Ko,(qty)	Subfamilies number (qty)	Characteristics of subfamilies	
					Sequences number (qty)	Invariant oligonucleotides R,(qty)
alcohol dehydrogenase	6	3	38	3	2	914
					2	50
					2	83
a- actin	12	27	38	2	6	47
					6	114

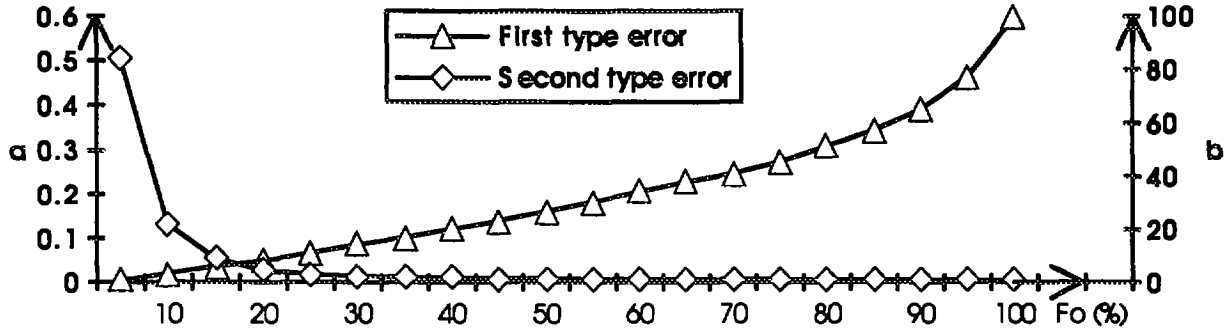

Figure 2. Errors at identification of nucleotide sequence for the families of protein-encoding isofunctional genes as depending on the threshold value *Fo*. a) First type error; b) Second type error.

were set up for the subfamilies. In all, a total of 563 significant vocabularies of invariant oligonucleotides l=8 bp in length has been built up for 172 gene families and 391 gene subfamilies, the average volume of a vocabulary being 156 entries.

Identification of nucleotide sequences by using vocabularies of invariant oligonucleotides

The simplest algorithm of identification is based upon the assumption that the more oligonucleotides, that are specific for a definite isofunctional family, is contained in an unknown sequence, the more the sequence is similar to the sequences of the family. The following threshold-setting rule is appropriate here to check whether the sequence X belongs to the family S:

$$Sequence\ X \begin{cases} \in S, if\ F(S) \geq F_0 \\ \notin S, if\ F(S) < F_0 \end{cases} \qquad (9)$$

Here $F(S)$ is the number of common oligonucleotides in the sequence X and in the vocabulary for the family S as a percentage of the total number of entries in this vocabulary. Fo is the threshold level of similarity such that, if not achieved, the sequence X does not belong to the family S. The threshold level Fo was set so as to risk the first and second type errors of identification. Estimates for Fo were obtained from analysis of independent (control) sequences that had not been used in building up the oligonucleotide vocabularies and had been selected by the well-known "jack knife" method (Arvesen, 1969).

For that purpose, from every family S_i (i=1,...,322) containing M_i sequences, one sequence s_i was canceled out at random for further use as a control. The other M-1 sequences made up the family S'_i for which a significant vocabulary W_i was built up. After analysis of all the families S'_i, the set of oligonucleotide vocabularies $\{W_i\}$ has been obtained, which was used for identification of the control sequences in line with rule (9). The sequences not involved into analysis made up the set $\{s_i\}$, to which we referred while assessing the accuracy of identification of nucleotide sequences with the oligonucleotide vocabularies on independent (control) data.

We considered a set of values for Fo at 0.05 increments: Fo=0.05, 0.1, 0.15,..., 0.95,1). At a fixed Fo, each of the control sequences of the set $\{s_i\}$ was related to one of the families $\{S_i\}$ in line with rule (9).

The first type error (false negative - FN) was evaluated as $FN=100(K-k)/K$. Here K is the total number of control sequences; k is the number of control sequences that have been related to the relevant families. If the sequence s_i was identified as being present at the family j ($i \neq j$) or as not being present at any family, we acknowledged an incorrect identification case.

The second type error (false positive - FP) was evaluated as $FP=\sum r_i / K$. Here r_i is the number of the sequences s_i identified as being present at the families S_j ($i \neq j$) (incorrect identification). This way of evaluation of the second type error allows one to estimate the average number of functional classes which a control sequence may be incorrectly identified as being present at. Ten iterations over the "jack knife" brought us to the means of the first and second type errors at different Fo.

The first type error at identification of the control sequences grows with Fo (Figure 2). Fo provides the least first type error at 5%≤Fo≤40%. In this interval, only from 0.4% to 12% of control sequences have not been identified as being present at the classes they ought to. The higher Fo, the less the second type error (Figure 2). This situation is because the higher the percentage of specific nucleotides observed at identification, the lower the probability of this oligonucleotide set being observed in the control sequence by chance. Fo provides the least second type error at 30%≤Fo≤100%. Within this interval, Fo takes account of just from 2.2 to 0.7 cases of incorrect

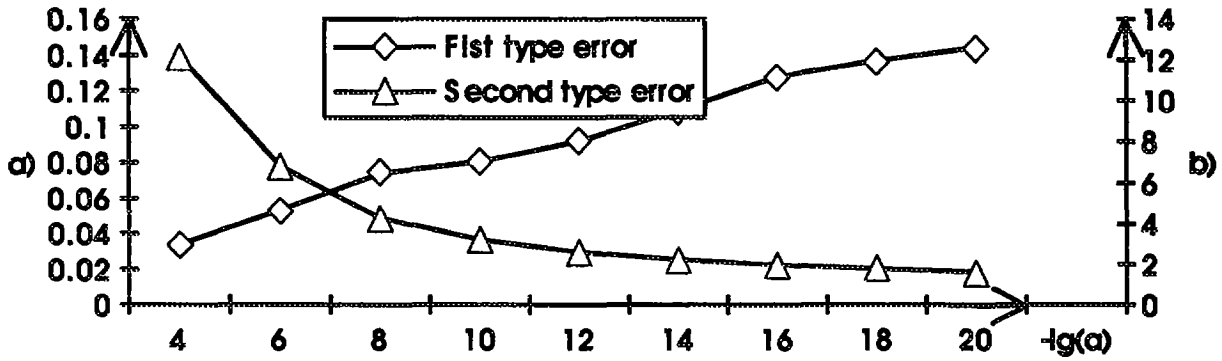

Figure 3. First and second type errors at the classification of nucleotide sequences by the statistical decision rule as depending on α. a) First type error b) Second type error

identification (i.e. the cases when sequences have been identified as being present at classes they do not belong to) per control sequence on the average.

By considering the intersection of two specified intervals, it is possible to evaluate the interval of variation for *Fo* which provides the concurrent minimization of the first and second type errors: 30%≤*Fo*≤40%. Here the first type error does not exceed 8.5-12%, and the second type error is from 2.3 through 1.7 incorrect identifications per control sequence. On the whole, the results obtained provide statistical evidence that the nucleotide sequences representing the coding gene regions can be identified reliably enough by using significant vocabularies of invariant oligonucleotides *l*=8 bp in length within the specified range of *Fo*.

Another method of identification with specific oligonucleotide vocabularies uses the following statistical decision rule:

$$Sequence\ X\begin{cases}\in S, if\ R(S)\geq K_0\\ \notin S, if\ R(S)<K_0\end{cases} \qquad (10)$$

Here *R(S)* is the number of oligonucleotides of the family *S* vocabulary, revealed in the sequence *X*, *Ko* is the upper limit as defined by Formulae (2)-(5), which is governed by the preset significance level α. That is why in this rule, α is an independent parameter which defines identification accuracy. To provide the minimization of the first and second type errors, the value of α was selected by the above scheme on the "jack knife" basis. Identification was performed at various values of α at power increments of 10^{-2} within the interval from 10^{-4} through 10^{-20}. As is seen (Figure 3), the values of α provide concurrent minimization of the first and second type errors if within an interval of $10^{-12}\leq\alpha\leq10^{-16}$. The first type error here does not exceed 10-14% and the second type error is from 2 to 3 incorrect identifications per control sequence. In general, the statistical decision rule at the most effective values of α provides much the same accuracy of recognition as the threshold-setting rule does at the most effective values of *Fo*.

Assessing the accuracy of identification of individual families of genes.

Here, 256 families out of 322 (79%) are noted for a zero first type error of identification (Fig 4a). It means that for any of these families the control sequence had been identified over 10 trials as being present at the relevant family.

Zero second type error was typical of 86 families (Figure 4b). It means that no control sequence had been identified as being present at an irrelevant family over 10 trials for these families. For 122 families, second type error was from 1 to 2 incorrect identifications per control sequence, and for 63 families from 2 to 4 incorrect identifications per control sequence (Figure 4b). For 26 families, both first and second type errors were zero. It means that their oligonucleotide vocabularies provide absolutely correct identification, i.e. control sequences have been identified as being present no otherwise than at the families they indeed belong to. Significant vocabularies of invariant nucleotides that provide accurate identification are exemplified in Table 3. Note that there are few entries in the vocabularies, which provides practical possibilities of using the revealed oligonucleotide vocabularies in experimental identification by hybridization of DNA

Table 3. Some of the families with zero first and second type errors at identification with the vocabularies of invariant oligonucleotides.

Family name	Oligonucleotides
carboxylase	ATGGATGC, AGCATGGA, AGGAAGCA, GTTGCACC, GGTTGCAC, GCATGGTT, CATGGTTG
NEF protein	ATTGGCAG, TTGGCAGA, TGGGTGGC, TGGCAGAA, GATTGGCA, GGTGGCAA, GGGTGGCA
somatostatin	TTCTTCTG, TTCTGGAA, TGCAAGAA, TCTTCTGG, TCCAGTGC, GGCTGCAA, GCTGCAAG, CAGTGCGC, CTTCTGGA, CTGCAACA, CGCAAAGC, CCTGTGCG

Figure 4. The distributions of the first (a) and second (b) type errors of identification over a sample of control sequences as determined by the statistical decision rule at the most effective value of $\alpha=10^{-14}$. As before, average estimates were obtained after 10 "jack knife" iterations.

sequences as belonging to the corresponding families.

Allowing for the least first type error, further perfection of the software should be guided at the development of methods for setting up oligonucleotide sequences that risk second type errors.

Evolutionary variables of the isofunctional gene family defining the volume of the vocabulary

The nucleotide sequences in the families studied display quite high homology (Table 1). Most of invariant oligonucleotides there must be located homologously at all sequences of the corresponding family. With due account of it, for each of 450 families (subfamilies) formed, we evaluated the parameter C, the value being called hereinafter the percentage of invariant oligonucleotides:

$$C=R/(L-l+1) \qquad (11)$$

C is the ratio of number R of observed invariant oligonucleotides of length l and their greatest possible number $(L-l+1)$ for a family of aligned sequences of average length L. The values of C for a range of the gene families studied are presented in column 8 of Table 1. For example, C=11% for the factor IX gene family and it is considerably higher for egg-laying hormone gene family (C=87%). Thus, we can observe eye-catching variation of C from family to family. At that, normalizing the observed number of oligonucleotides R to the average length L of a family sequence in line with (11) brings up such calculus of C that sequence length no longer affects the observed number of invariant oligonucleotides.

It is of interest to determine the factors accounting for the range of C. For this purpose, consider a model that allowed us to estimate expectations C^* for the family of aligned nucleotide sequences of length L. It is now declared that a position of alignment is a variable position if there are more than one variant of the nucleotide at that position (Figure 5). The expected number of invariant oligonucleotides of length l may be assessed as by Zharkikh and Rzhetsky, 1993, as follows. The number of permutations of V variable positions at a sequence of length L is C_L^V. The number of permutations of V variable positions at which a specified oligonucleotide of length l is unaffected is C_{L-l}^V. Then the probability of the specified oligonucleotide of length l not containing any variable position is C_{L-l}^V/C_L^V. By multiplying this probability by $(L-l+1)$, which is the total number of oligonucleotides in the sequence of length L, we obtain the expected number of invariant oligonucleotides:

$$K(L,V,l)=(L-l+1)C_{L-l}^V/C_L^V \sim (L-l+1)(1-\frac{V}{L})^l \qquad (12)$$

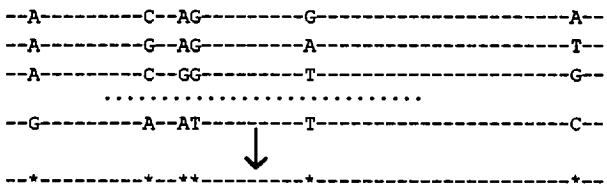

Figure 5. A set of aligned nucleotide sequences. Variable positions are asterisked.

Equation (12) provides true estimates given that there are V variable positions evenly distributed along the sequence. After rearrangements, we derive from (12)

$$C^*(L,V,l)=\frac{K(L,V,l)}{L-l+1} \sim (1-\frac{V}{L})^l=(1-Q)^l \qquad (13)$$

where $Q=V/L$ is the percentage of variable positions in the alignment.

C^*, which is called the expected percentage of invariant oligonucleotides, is the ratio of the expected number $K(L,V,l)$ of invariant oligonucleotides of length l and the greatest possible number $(L-l+1)$ of invariant oligonucleotide of such length for any family of aligned sequences of length L each. By using (13) one can

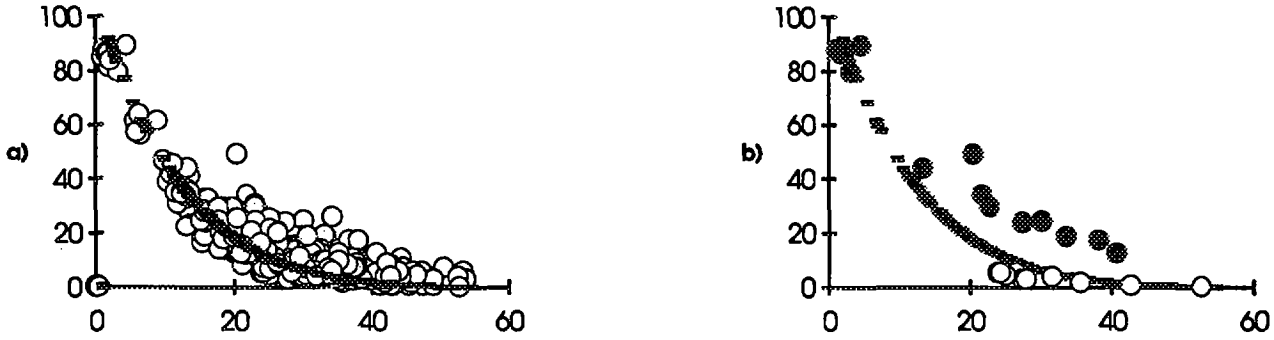

Figure 6. Comparison of the theoretical (solid line) and observed dependencies of the percentage of invariant oligonucleotides on the percentage of variable positions; the vertical axis is C^*; the horizontal axis is Q. a) Comparison for the entire pool of gene families (each family is symbolized by "o"). b) Comparison for the families with strongly (•) and weakly (o) uneven distributions of variable positions.

calculate C^* for each family on the basis of the average length, L, of a family sequence, oligonucleotide length l and the number, V, of variable positions.

The observed Q's for a range of gene families are presented in the last but one column of Table 1. For example, for the egg-laying hormone gene family, Q=1.6%, whereas for the collagen gene family it is considerably higher (Q=53%). Hence, there is considerable variation of Q amongst the families.

The theoretical dependence C^*/Q calculated from (13) is presented in Figure 6a. The observed dependence $C(Q)$ for all the families is there, too. As is seen, in general the observed $C(Q)$ is well consistent with the theoretical dependence. It means that the percentage of invariant oligonucleotides in the family vocabulary is largely governed by the percentage of variable positions. It is seen, however, that strong variation of C's is observed within the range $Q \geq 20$. Table 1 presents gene families, that at close Q's have essentially different C's. For example, for the family of phosphoglycerate kinase, C=49% at Q = 21. However, for the alkaline phosphatase families, the C is much different (12) at close Q=24. It means that there must be some additional factor(s) that strongly determines the volume of the oligonucleotide vocabulary.

Unevenness in the distribution of variable positions along the sequences appears to be one. A qualitative illustration of its effect is presented in Figure 7. Let l be a fixed length of an oligonucleotide. One variable position (*) accounts for the elimination of l overlapping oligonucleotides from their full register. In Figure 8, 8 oligonucleotides are eliminated at l=8. If there is another variable position $I \geq l$ bp away from the first, l more oligonucleotides are eliminated. The total of oligonucleotides eliminated in this case is $2xl$ (as in Figure 7a, when the second variable position is 17 bp away from the first).

If the second variable position (+) is $I < l$ bp away from the first (Figure 7b), the first variable position accounts for the elimination of l oligonucleotides (here l=8), the second, as oligonucleotides are overlapping, accounts for only $l - I$ oligonucleotides (here l-I=4). A third variable

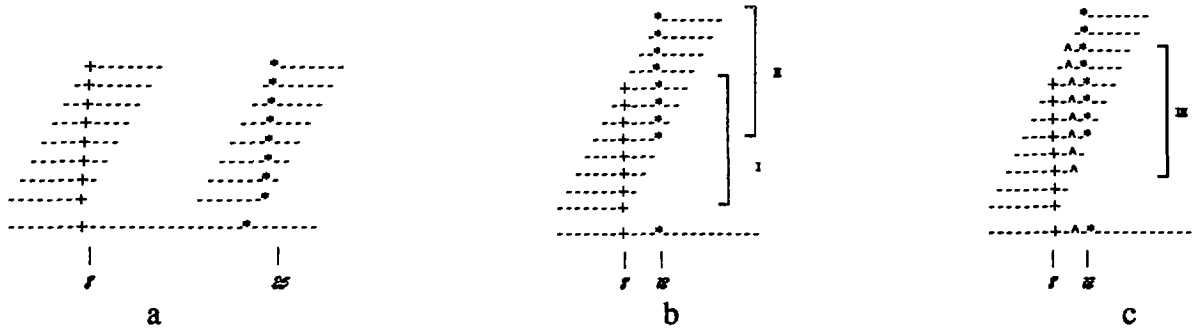

Figure 7. Uneven distribution of variable positions as affecting the number of eliminated oligonucleotides (l=8). a) The distance between two variable positions (* and +) exceeds oligonucleotide length. b) The distance between two variable positions (* and +) less than oligonucleotide length. I - invariant oligonucleotides eliminated due to the first variable position. II - invariant oligonucleotides eliminated due to the second variable position. c) the third variable position (^) is located between the two closely set variable positions. III - invariant oligonucleotides eliminated due to the third variable position.

position (^) occurring between other two positions that are less than 8 bp apart (Figure 7c) causes no additional elimination of invariant oligonucleotides. In such situation, the set of oligonucleotides with the third position (^) is shared between the sets of oligonucleotides that contain the first and second positions.

So, when variable positions are as close as above declared, a kind of "interference" takes place in the stretch between them: the higher the number of regions with small distances between variable positions, the less, *ceteris paribus*, the number of invariant oligonucleotides liable to be eliminated from the full register. On the way towards Expressions (12)-(13), the assumption that variable positions are independently distributed along the sequence in that the "interference" does not take place, was quite of essence. It is apparent that the observed discrepancies between C and C^*, as well as high variation in C at fixed Q's are accounted for exactly by the unevenness of the distribution of variable positions and by "interference".

Unevenness can be assessed as $G = S_x/X$, where X is the average distance between variable positions, S_x is the standard deviation of X. The higher G, the higher unevenness. G was evaluated for all gene families under study (see the last column of Table 1). For example for the factor IX gene family, G = 0.89, from whence evenness was assumed. While for the gene family of phosphoglycerate kinase, G=2.8, which gave out sure unevenness. On the whole, for the gene families studied, G varies within one order of magnitude (Table 1).

In Figure 6b we compare the theoretical dependence of the percentage of invariant oligonucleotides, C, on the percentage of variable positions, V, against the observed dependencies between these values as for two subgroups of the families. One of the subgroups consisted of 5% of the families with the strongest unevenness, whereas the other consisted of 5% of the families with the weakest unevenness. As is seen, for the families with the greatest G's, the $C^*(Q)$ dots fall over the theoretical curve, and for the families with the least G's, under. This convincingly illustrates that the parameter of unevenness, G, essentially accounts for the difference between the observed and expected numbers of invariant oligonucleotides.

Acknowledgments

This work was supported by the Russian Foundation of Fundamental Investigations (grant N 94-04-13241-a), Russian National Human Genome Project, Russian Ministry of Sciences and Technique Politics and USA Department of Energy. The authors are thankful to V.Filonenko for translation of this manuscript from Russian into English.

References

Arvesen, J. 1969. Jacknifing U-statistics. Ann. Math. Statist. 40: 2076-2100.

Davies, K.E. ed. 1986. *Human genetic diseases: a practical approach.* IRL Press Limited: Oxford.

Hoheisel, J. D. 1994. Application of hybridization techniques to genome mapping and sequencing. *Trends in Genetics.* 10(3): 79-83.

Khrapko, K. R., Lysov, Yu. P., Khorlyn, A. A., Shick, V. V., Florentyev, V. A., Mirzabekov, A. D. 1989. An oligonucleotide hybridization approach to DNA sequencing. *FEBS Lett.* 256: 118-122.

Khrapko, K. R., Khorlin, A. A., Ivanov, I. B., Chernov, K. K., Lysov, Yu. P., Vasilenko, S. K., Florentyev, V. A., Mirzabekov, A. D. 1991. Hybridization of DNA with gel immobilized oligonucleotides: convenient method of single substitution revealing. *Molekuliarnaya Biologiya* 25(3): 718- 730. (in Russ.)

Kuznetsova, S. A., Kanevsky, I. E., Florentyev, V. A., Mirzabekov, A. D., Shabarova, Z. A. 1994. Sequencing by hybridization to oligonucleotides immobilized in gel. Chemical ligation as further advantage of the technique. *Molekuliarnaya Biologiya* 28(2): 290-299 (in Russ).

Saiki, R. K., Bugawan, T. L., Horn, G.T., Mullis, K. B. and Erlich, H.A. 1986. Analysis of enzymatically amplified beta-globin and HLA-DQalpha DNA with allele-specific oligonucleotide probes. *Nature* 324: 153-166.

Shindyalov, I. N., Kolchanov N. A. 1993. A computer system for the analysis of molecular evolution in isofunctional gene families. *International Journal of Genome Research* 1(2): 129-148.

Solovyev V. V., Seledsov I. A. 1993. A new approach to the phylogenetic trees construction based on the analysis of the relatively conservative regions of the nucleotide and amino acid sequences. *International Journal of Genome Research* 1(3): 177-185.

Strezoska, Z., Paunesku, T., Radosavljevic, D., Labat, I., Drmanac, R., Crkvenjakov, R. 1991. DNA sequencing by hybridization: 100 bases read by a non-gel-based method. *Proc. Natl. Acad. Sci. USA* 88: 10089-10093.

Suggs., S. V., Wallace, R. B., Hirose, T., Kawashima E. H., and Itakura, K. 1981. Use of synthetic oligonucleotides as hybridization probes: isolation of cloned cDNA sequences for human beta2-microglobulin. *Proc. Natl. Acad. Sci. USA* 78: 6613-6617.

Zharkikh, A. A., Rzhetsky, A. Y. 1993 . Quick assessment of similarity of two sequences by comparison of their L-tuple frequencies. *Biosystems* 30: 93-112.