selection testing, human cancer genes,
ATM, RECQL

Krzysztof CYRAN[*,**], Joanna POLAŃSKA[***],
Marek KIMMEL[*,***]

TESTING FOR SIGNATURES OF NATURAL SELECTION AT MOLECULAR GENES LEVEL

The paper presents the methodology used for detecting the signatures of natural selection at the molecular level from single nucleotide polymorphism data. The results obtained from widely used approach, based on statistical testing departures from neutral evolution model, can be obscured by the presence of alternative hypotheses generating the similar to natural selection results of the tests. These hypotheses include population growth and geographic substructure. Especially for human population these alternatives are of non-negligible importance. In the paper we show how to deal with this problem, both by the analysis of a battery of statistical tests giving indication about the age of the predominant mutations, and by application of non conventional null hypotheses that assume different population scenarios. Since the critical values of the tests are known only for panmicting, constant size population, the second approach demands the intensive computer simulations of coalescence process to obtain analogous critical values for different scenarios used as a null. The methodology with the problem of detecting signatures of natural selection in four genes implicated in human familial cancers has been illustrated:

1. INTRODUCTION

There exist two general types of tests of natural selection on the molecular level. The first type can be applied when the data consists of entire or partial coding sequences of a gene. Then the comparison of frequencies of silent substitutions at the third codon position to the frequencies of substitutions on the first and second position, provides a handle to measure selective pressure. This approach was used in study leading to detection of perhaps the most spectacular example of natural selection found in the ASPM locus, a major contributor to brain size regulation in primates [6, 18].

In many cases, however we have to deal with another type of data, which consists of sequences that are not only non-coding, but also composed of nucleotides located at a considerable distance from each other. In such cases, a model for neutral evolution of the sequence has to be determined and then its predictions compared to data. Usually, this

[*] Department of Statistics, Rice University, Houston TX, USA
[**] Institute of Computer Science, Silesian University of Technology, Gliwice, Poland
[***] Institute of Automation, Silesian University of Technology, Gliwice, Poland

model is some modification of the Wright-Fisher model of genetic drift with mutation [9, 10]. Examples include population substructure and past change in population size [13]. Therefore, one common way to deal with this problem is to frequently apply a number of tests, each one sensitive to different combination of factors, and compare the results. The substructure for example can be approached by considering data from different subpopulations separately or by comparison of the test results among loci. Another approach presented in the paper is based on the formulation of null hypotheses assuming population substructure. Single Nucleotide Polymorphisms (SNP) data taken from the intronic regions of a target gene provides an example of data of the second type. They form haplotypes, which can be used as tools to investigate the genetic diversity and possible disease association of the target gene. We considered in a series of previous papers SNP haplotypes at four genes implicated in human familial cancers. The first locus analysed is ataxia telangiectasia (ATM) [1, 2]. The ATM gene product is a member of a family of large proteins implicated in the regulation of the cell cycle and response to DNA damage [16]. The other three genes include: human helicase RECQL, Bloom's syndrome (BLM) and Werner's syndrome (WRN) [15]. The products of these three genes are DNA helicases, enzymes involved in various types of DNA repair, including mismatch repair, nucleotide excision repair, and direct repair. A number of interesting facts about these genes were determined, however, the question of selection signatures was not addressed before. In addition to standard tests, in this study we use also simulation charts being a sort of graphical visualisation of departures from the Ewens Sampling Formula.

2. METHODOLOGY

Each analysis of SNP data leading to the detection of natural selection operating at some loci, when applied to human population, has to take into consideration the alternative departures from neutrality that can produce data resulting in similar test outcomes. These alternatives feasible from the point of view of human population evolution are population growth and geographic substructure with migration. Here we show how to deal with this problem, by the analysis of a battery of statistical tests giving indication about the age of the predominant mutations, and how this information can be used to exclude not desirable alternatives. The tests which give the indication about the age of alleles, being in excess compared to the amount predicted under neutral evolution model, are based on the difference between different estimates of composite parameter $\theta = 4N\mu$ (N indicates the effective population size and μ is the mutation rate per nucleotide per generation). Such tests are Fu's tests belonging to the class $F'(r, r')$ [8]:

$$F'(r,r') = \frac{L'(r) - L'(r')}{\sqrt{Var\left[L'(r) - L'(r')\right]}}, \qquad (1)$$

where L' are estimates of composite parameter θ in the form of linear functions of the η_i (the numbers of segregating sites of type i, where $i = 1, 2, \ldots, \lfloor n/2 \rfloor$ and n is the sample

size). The parameter of function *L'* denotes more (for larger values) or less (for smaller values) substantial influence of rare alleles on the estimation of θ. Therefore, $\hat{\theta}_\pi = L'(0)$ is less influenced by rare alleles than $\hat{\theta}_W = L'(1)$. The defined above class covers many known tests like: Tajima test *T* (for uniformity, we follow the nomenclature of Fu [8, 17] and some other papers, although originally Tajima's test was named *D*), Fu and Li's test *D** or Fu and Li's test *F**. The Tajima *T* test, which is the most widely used neutrality test [12], is equivalent to *F'*(0,1) and is defined as the normalised difference between the estimates of composite parameter $\theta = 4N\mu$ based on the average genetic distance $\hat{\theta}_\pi$ and the number of segregating sites:

$$T = \frac{\hat{\theta}_\pi - \hat{\theta}_W}{\sqrt{Var\left(\hat{\theta}_\pi - \hat{\theta}_W\right)}}. \qquad (2)$$

Others tests of *F'*(*r*, *r'*) class include Fu and Li's test *D** (*D** = *F'*(1,∞) and therefore the test is sensitive to existence of very rare alleles) and Fu and Li's test *F**. Since *F** = *F'*(0, ∞) it should have the power for detecting the excess of very rare alleles, presumably with greater power than *D** because of a more extreme value of the first parameter in function *F'*. These tests are defined as [7]:

$$D^* = \frac{\frac{n}{n-1}\eta - \eta_s \sum_{i=1}^{n-1} \frac{1}{i}}{\sqrt{u_{D^*}\eta + v_{D^*}\eta^2}}, \qquad F^* = \frac{\hat{\theta}_\pi - \frac{n-1}{n}\eta_s}{\sqrt{u_{F^*}\eta + v_{F^*}\eta^2}}, \qquad (3)$$

where η is the total number of mutations that occurred in the entire genealogy of *n* genes, and η_s is the number of singletons, *i.e.* nucleotides that appear only once at the site among the sequences in the sample. For mathematical definitions of coefficients u_{D^*}, v_{D^*}, u_{F^*} and v_{F^*} (being complicated functions of the parameter *n* only) see Fu and Li [7].

Another category of tests is based on the estimates of probabilities of having no more or no less than the observed number *k* of haplotypes in a sample of *n* sequences, assuming neutrality and lack of intra-locus recombination. Into this category fall: Strobeck's test *S* (the estimate of the probability of having no more haplotypes in a sample) and Fu's test F_s. The Strobeck's test *S* is given by:

$$S = \sum_{i=1}^{k} \frac{\left|S_n^i\right| \hat{\theta}_\pi^i}{S_n\left(\hat{\theta}_\pi\right)} \qquad (4)$$

where: $S_n(\hat{\theta}_\pi)$ denotes the generating function of the Stirling numbers of the first kind S_n^i, *i.e.* $S_n(\hat{\theta}_\pi) = \sum_{i=0}^{n} S_n^i \hat{\theta}_\pi^i = \hat{\theta}_\pi(\hat{\theta}_\pi + 1)\ldots(\hat{\theta}_\pi + n - 1)$. Fu's test F_s is given by:

$$F_s = \ln\left(\frac{S'}{1-S'}\right), \tag{5}$$

where S' is the estimate of the probability of having no less than observed number k of haplotypes in a sample of n sequences. Therefore (compare with (4) for similarities) it is given by:

$$S' = \sum_{i=k}^{n} \frac{|S_i|\hat{\theta}_\pi^i}{S_n\left(\hat{\theta}_\pi\right)}. \tag{6}$$

In the framework of the infinite allele model (IAM), the SNP haplotypes are treated as new variants (mutants) of a SNP sequence. Ewens Sampling Formula, derived under neutrality and no recombination, provides expected frequencies of haplotypes existing in a given number of copies [9]. Therefore, it serves as a convenient reference to test deviations from neutrality. It is used, for example, in the Strobeck's test (see above). However, it is even more convenient to use coalescent simulations based on the IAM, to compute a large sample of simulated distributions of variants. The value of composite parameter θ is estimated from the haplotype sample, using the IAM-based expression for the total number K of variants in the sample of n sequences:

$$K(\theta) = \sum_{i-1}^{n} \frac{\theta}{\theta + i - 1} \tag{7}$$

and comparing it to the observed number of different haplotypes. Simulated distributions of variants are compared to the observed frequencies. Technically, to facilitate visual comparison, empirical and simulated cumulative counts $A(j)$ of haplotype variants existing in no more than j copies in the sample of n sequences (j = 1, ..., n) are compared. In addition, both the horizontal axis (the number j of copies of a variant) and the vertical axis (cumulative count $A(j)$ of variants existing in j copies) are standardised to the unit interval, by dividing by n and K, respectively. Resulting graphs allow a visual comparison of the empirical distribution of variants (thick line) with multiple simulated distributions (thin lines), as presented in Fig. 1 for actual SNP data.

If the goal is eventually to find the type of selection, at first one should exclude Kelly's Z_{nS} test, as it produces similar, inflated, patterns both for selective sweeps with recombination and for balancing selection. However, it is valuable to apply the Z_{nS} test after one of these possibilities has been excluded based on results from tests (2) - (5). It is so because this test is reported to have a big power, and can verify previously obtained results. It is defined as the average (over all pairs i, j of K segregating sites) of the squared correlation of allelic identity between sites i and j [11]:

$$Z_{nS} = \frac{2}{K(K-1)} \sum_{i=1}^{K-1} \sum_{j=i+1}^{K} \delta_{ij.} \tag{8}$$

The squared correlation of allelic identity δ_{ij} is a standardised (that is ranging from 0 to 1) measure of linkage disequilibrium D_{ij} between loci *i* and *j*. It is given by:

$$\delta_{ij} = \frac{D_{ij}^2}{p_i(1-p_i)p_j(1-p_j)}. \qquad (9)$$

In above formula $D_{ij} = p_{ij} - p_i p_j$, where p_i and p_j are frequencies of mutant alleles at loci *i* and *j* respectively, whereas p_{ij} is the frequency of sequences that have mutant alleles at both loci.

The last tests we present here are Wall's tests *B* and *Q* [17]. Statistic *B* is the normalised number *B'* of pairs of adjacent congruent (*i.e.* inducing identical partitions of the set of haplotypes) segregating sites. To be normalised, *B'* is divided by the total number $(K-1)$ of pairs of adjacent segregating sites: $B = B'/(K-1)$. If we indicate by A the set of all distinct partitions induced by pairs of adjacent congruent segregating sites, then the statistic $Q = [B + card(A)]/K$, where $card(A)$ is the number of elements in *A* and the power of a test becomes less sensitive to the recombination, because of compensation of the decrease of *B* by an increase of *card* (A).

The careful analysis of the mentioned battery of tests if applied to many loci and for many subpopulations can give the answer about the presence of selection at some of them. Theoretically, population growth and population substructure effects should be identical (or in the presence of recombination, similar) for all loci. Any large difference in test outcomes among loci is a signal that some specific to some loci reason is probably non-negligible cause of detected departures. Also analysis of relatively genetically pure subpopulation can reveal that the cause of departure from neutrality is not the substructure (since in such subpopulation it is not of the main importance). Yet in practice it is not easy to obtain a sample from genetically pure population since admixtures accumulated over long time are of different intensity in main human subpopulations [3-5].

Because of mentioned reasons it may be helpful to employ more sophisticated null hypotheses. Certainly they should assume neutrality (which is subject to be rejected by the test result) but on contrary to standard null hypotheses they can incorporate more feasible population models. The degree to which they can imitate the real history of human population depends on our knowledge about this history (still very incomplete in long term) but they always should be formulated to be conservative with respect to feasible population history scenarios (in order to prevent too many false positives). The exact meaning of being conservative in this aspect is dependent on the actual data. Therefore it is always desirable to perform the battery of mentioned above tests with standard null hypotheses and infer based on them whether departure from neutral model is in direction of excess of old or young mutations. The excess of young mutations is characteristic for positive selective sweep or for slightly deleterious mutations, whereas the excess of old mutations is observed in loci under balancing selection pressure. Since the population expansion is also the cause of many young mutations, therefore null hypothesis assuming growth would be more conservative than standard in search for selective sweep, but less conservative than standard in search for balancing selection. On the other hand the effect of population substructure shifts the excess of alleles in opposite direction as compared to population growth.

3. APPLICATION TO SNP DATA

As an application of presented in above section methodology, we analysed 45 intron SNPs in the four helicases: ATM, BLM, WRN, and RECQL. The detailed information on primer sequences, PCR conditions and product size for each of the polymorphic sites, as well as the ASO sequences and wash conditions for each SNP variant can be found in [1, 15]. Blood samples were collected from the individuals, residents of Houston, TX, from four major ethnic groups: Caucasians, Asians, Hispanics, and African-Americans. Haplotypes were inferred and their frequencies were estimated by using the EM algorithm [14]. The results of tests T, D^*, F^*, F_s, S and Z_{nS} are given in a Table 1.

Table 1. Summary of the results of nonneutrality tests: T, D^*, F^*, F_s, S and Z_{nS}.

		T	D^*	F^*	F_s	S	Z_{nS}
ATM	Global	3.88 ***	1.46 ?	2.87 **	2.42	0.13 NS	0.39 *
	AfAm	2.42 *	1.50 ?	2.10 *	1.94	0.20 NS	0.29 NS
	Caucasian	3.48 ***	1.54 *	2.60 **	10.88	0.00 ***	0.47 *
	Asian	2.55 *	-0.07 NS	0.96 NS	2.81	0.11 ?	0.49 *
	Hispanic	3.20 **	1.54 *	2.47 **	2.17	0.17 NS	0.45 *
RecQL	Global	3.83 ***	1.30 ?	2.70 **	-14	1 NS	0.32 ?
	AfAm	2.83 **	0.69 NS	1.68 ?	-3.57	0.99 NS	0.24 NS
	Caucasian	3.10 **	1.40 ?	2.30 **	-1.05[d]	0.82 NS	0.36 ?
	Asian	2.65 *	0.71 NS	1.52 ?	1.89	0.23 NS	0.52 *
	Hispanic	2.93 **	1.40 ?	2.23 **	-8.12	1.00 NS	0.32 ?
WRN	Global	1.57 NS	1.35 ?	1.72 ?	-26	1 NS	0.08 NS
	AfAm	0.79 NS	-0.15 NS	0.21 NS	-13.58	1.00 NS	0.06 NS
	Caucasian	1.26 NS	1.45 ?	1.58 NS	-14.79	1.00 NS	0.10 NS
	Asian	1.36 NS	1.33 ?	1.47 ?	-1.70	0.92 NS	0.18 NS
	Hispanic	1.10 NS	-0.56 NS	0.05 NS	-5.83	1.00 NS	0.12 NS
BLM	Global	2.97 **	1.12 NS	2.14 *	-30	1 NS	0.14 NS
	AfAm	2.06 ?	1.23 NS	1.72 ?	-11.69	1.00 NS	0.12 NS
	Caucasian	2.50 *	1.23 NS	1.90 *	-11.26	1.00 NS	0.18 NS
	Asian	1.78 ?	1.28 NS	1.58 ?	-4.85	1.00 NS	0.17 NS
	Hispanic	1.87 ?	1.23 NS	1.65 ?	-7.80	1.00 NS	0.15 NS

The shaded region indicates that the outcomes of tests are significant or close to significance for almost all subpopulations. This is the case for ATM and RecQL genes as detected by $F(r,r')$ tests. Observe that the test values are all positive, reflecting the excess (with respect to neutral evolution model predictions) of the old mutation in a sample.

Diagonally shaded region emphasises significant or insignificant but close to significant outcomes of S test and positive values of F_s test. For ATM shaded and diagonally shaded region overlap indicating evidence for departure from neutrality in the direction of old alleles. For RecQL the lack of overlap can be explained by relatively high rate of recombination, which radically decreases the power of neutrality tests based on the number of haplotypes (like S and F_s). Additionally, for Wall's B and Q tests, we applied three different null hypotheses: H_{00} (panmictic population with size constant in time), H_{01} (panmictic population with size increasing 10 times over the period of 5000 human generations, according to an exponential function) and H_{02} (substructured population composed of 4 demes with between-deme migration rate $mN = 10$, with total size increasing 10 times over the period of 5000 human generations, according to an exponential function). Each of the above assumes selective neutrality.

Since we already determined the excess of old mutations and only looked for the plausible explanation of them, we could deduce that hypotheses H_{01} and H_{02} were less conservative than H_{00}, although they were still conservative (in the aspect of preserving the excess of old mutations), considering the feasible scenarios of human population history. The reason for this is that actual size increase of human population was presumably larger than 10-fold growth over 5000 generations, as assumed in H_{01}. Such larger growth makes the hypothesis H_{01} conservative if the direction of departures from neutrality is towards excess of old mutations. At this stage of analysis however, based on the outcomes of tests (2)-(5) we already knew that this is the case and therefore when formulating these hypotheses we considered departures from neutrality only in this direction. Since hypothesis H_{02} is always more conservative in the aspect of preserving the excess of old mutations than H_{01}, so if H_{01} is conservative so must be also H_{02}. The Wall's tests results for modified null hypotheses have the same general pattern like tests (2) - (5) and test (8). The graphical depiction of the departure from neutrality for ATM and lack of departure for WRN is given in Fig. 1 for African American population.

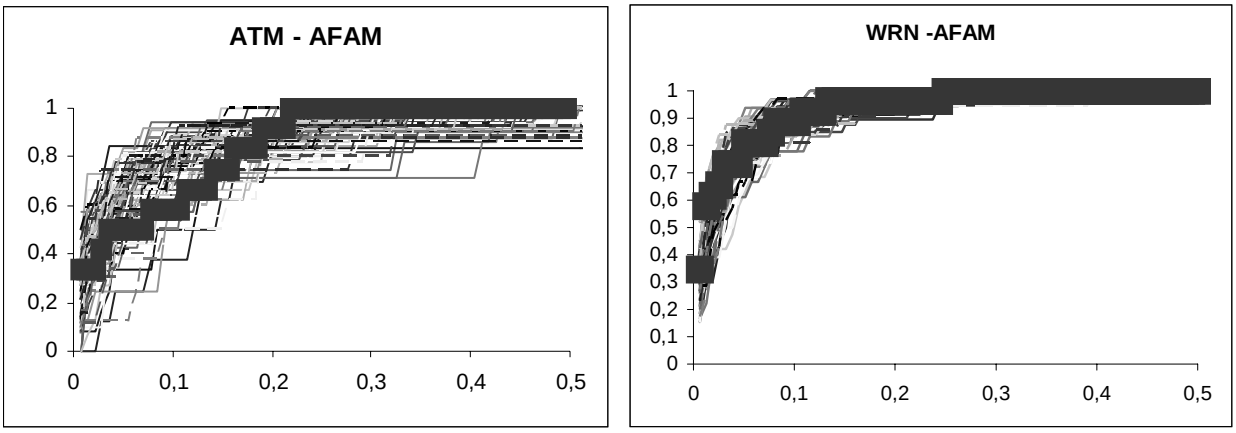

Fig. 1. Graphical depiction of the nonneutrality at ATM locus obtained from simulations

4. CONCLUSIONS

According to $F'(r, r')$ tests such as T, F^* and D^* the loci analysed fall into two categories, A (ATM, RecQL) and B (WRN, BLM). Category A is characterised by the positive and, on contrary to B, statistically significant outcomes. According to tests based on the number of haplotypes in a sample, the results for ATM are different from the results for three other loci. Kelly's Z_{nS} test outcomes are essentially the same as those of the $F'(r, r')$ tests. The same partition is observed according to Wall's B and Q tests, though almost no outcome is significant in these tests against null hypothesis H_{00}. Results of these tests against the slightly less conservative null hypotheses H_{01} and H_{02} are significant for genes ATM and RecQL but still not significant for any population at BLM and WRN loci. Using distributions of haplotype variants simulated from the IAM, we constructed cumulative graphs. These results corroborate the outcome of the Strobeck's test.

Nielsen [13] suggests being conservative in concluding about detection of selection based on tests using only haplotype data, because other alternatives lead to similar results. The main alternative is that of population growth. Further analysis of obtained results shows, however, that these concerns, which are especially important in the case of selective sweeps leading to an excess of young mutations [8], are not applicable directly to this study, with samples displaying excess of old mutations. To decide whether tests reflect a natural selection or population substructure is more complicated, but can be addressed by parallel analysis of more than one locus. Since the geographical subdivision is identical for different loci, so should be the results of tests, if the loci are neutral. Even in the presence of recombination among loci, the test results should remain similar, since despite the different genealogies, expectations of parameters used in tests under the same neutral model should be the same for all loci. However, this is not the case. WRN and BLM unlike ATM and RecQL, are not statistically significant. So the overall conclusion is that the ATM and RecQL are subject to special type of selection resulting in samples with a strong excess of old mutations. Selection may operate in a linked coding region, with genetic hitchhiking being responsible for transmission of departures to analysed intronic regions.

BIBLIOGRAPHY

[1] BONNEN, P.E., M. D. STORY, C. L. ASHORN, T. A. BUCHHOLZ, M. M. WEIL, D. L. NELSON. 2000. Haplotypes at ATM identify coding-sequence variation and indicate a region of extensive linkage disequilibrium. Am J Hum Genet 67: 1437-1451.

[2] BONNEN, P.E., P. J. WANG, M. KIMMEL, R. CHAKRABORTY, D. L. NELSON. 2002. Haplotype and linkage disequilibrium architecture for human cancer-associated genes. Genome Res 12: 1846-1853.

[3] BUDOWLE, B., R. CHAKRABORTY. 2001. Population variation at the CODIS core short tandem repeat loci in Europeans. Legal Medicine 3: 29-33.

[4] BUDOWLE, B., B. SHEA, S. NIEZGODA, R. CHAKRABORTY. 2001. CODIS STR Loci Data from 41 Sample Populations. Journal of Forensic Sciences 5: 453-489.

[5] CHAKRABORTY, R. 1986. Gene Admixture in Human Populations: Models and Predictions. Yearbook of Physical Anthropology 29: 1-43.

[6] EVANS, P. D., J. R. ANDERSON, E. J. VALLENDER, S. L. GILBERT, Ch. M. MALCOM, S. DORUS, and B. T. LAHN. 2004. Adaptive evolution of ASPM, a major determinant of cerebral cortical size in humans, Human Molecular Genetics, in press.
[7] FU, Y. X. and W. H. Li. 1993. Statistical Tests of Neutrality of Mutations, Genetics 133: 693-709.
[8] FU, Y. X. 1997. Statistical Tests of Neutrality of Mutations Against Population Growth, Hitchhiking and Background Selection, Genetics 147: 915-925.
[9] HARTL, D. L. and A. G. CLARK. 1997. Principles of Population Genetics. Sinauer Assoc., Sunderland, MA.
[10] JOBLING, M.A., M. E. HURLES, and C. TYLER-SMITH. 2004. Human Evolutionary Genetics: origins, peoples & disease, Garland Science, New Delhi, India.
[11] KELLY, J. K. 1997. A test of Neutrality Based on Interlocus Associations, Genetics 146: 1197-1206.
[12] MCVEAN, G. 2002. Natural Selection, Printed Materials of Univ. Oxford, Dept. Stat.:1-25.
[13] NIELSEN, R. 2001. Statistical tests of selective neutrality in the age of genomics, Heredity 86: 641-647.
[14] POLAŃSKA, J. 2003. The EM algorithm and its implementation for the estimation of the frequencies of SNP-haplotypes. Int. J. Appl. Math. Comput. Sci. 13: 419-429.
[15] TRIKKA, D., Z. FANG, A. RENWICK, S. H. JONES, R. CHAKRABORTY, M. KIMMEL, D. L. NELSON. 2002. Complex SNP-based haplotypes in three human helicases: implications for cancer association studies. Genome Res 12: 627-639.
[16] VOROCHEVSKY, I., L. LUO, A. LINDBLOM, M. NEGRINI, A. D. WEBSTER, C. M. CROCE, and L. HAMMARSTROM. 1996. ATM mutations in cancer families. Cancer Res. 56: 4130-4133.
[17] WALL, J. D. 1999. Recombination and the power of statistical tests of neutrality, Genet. Res. 74: 65-79.
[18] ZHANG, J. 2003. Evolution of the Human ASPM Gene, a Major Determinant of Brain Size. Genet. 165: 2063-2070.

This work was supported by a grant CA 75432 from the National Cancer Institute of the NIH and by a grant 4T11F 01824 from Polish State Committee for Scientific Research (KBN).