Pattern Building Methods in Genetic Data Processing

Tatyana Volkova, Elena Furta, Olga Dmitrieva, Irina Shabalina

***Abstract*—This work is a part of long-term investigations carried out by the Laboratory of Molecular Genetics of Constitutive Immunity at Petrozavodsk State University. The paper describes development of approaches to genetic information analysis as applied to septic shock sensitivity investigation on basis of a mice model. The septic shock can be modeled by the injection of tumor necrosis factor (TNF). The paper describes two methods of finding markers combinations (patterns) affecting resistance to TNF. The developed algorithms were implemented and applied to the input data. The results showed the significant correspondence between some chromosome markers and resistance to TNF.**

***Index Terms*—Genetic Analysis, Mathematical Methods of Diagnostics, Statistical Methods**

I. Introduction

DIFFERENT species of animals have various levels of sensitivity to septic shock. A mice model is used for the investigation of human sensitivity. Previous investigations showed that septic shock can be modeled by the injection of the synthetically obtained tumor necrosis factor (TNF) [1], [2].

Two lines of mice were used in the investigation. The mice belonging to *B6* line have sensitivity to TNF similar to a human one, while the absolutely resistant to TNF mice belong to *Msm* line.

The genetic analysis method can be applied to search the gene responsible for TNF sensitivity using two mice lines with opposite TNF sensitivity. This method considers the outcrossing of different lines of mice and the analysis of a phenotype and a genotype of the rising generation. The analysis is performed by comparison of the observed phenotype and the determined genotype for the rising generation objects.

Manuscript received May 15, 2014; accepted June 5, 2014. Date of online publication: June 30, 2014.

Authors are grateful for the Strategic Development Program of Petrozavodsk State University that provided required support for this research.

T. Volkova is with the Laboratory of Molecular Genetics of Constitutive Immunity at Petrozavodsk State University (e-mail: volkovato@yandex.ru).

E. Furta is with the Faculty of Mathematics at Petrozavodsk State University (e-mail: lena-furta@rambler.ru).

O. Dmitrieva is with the Faculty of Mathematics at Petrozavodsk State University (e-mail: dmitr_eva@mail.ru).

I. Shabalina is with the Chair of Applied Mathematics and Cybernetics, Faculty of Mathematics at Petrozavodsk State University (e-mail: i_shabalina@petrsu.ru).

There are two types of phenotypes: a resistant type (a mouse survives after TNF injection) and a sensitive type (a mouse dies after TNF injection).

The genotype can be represented by three types: “A” – a resistant genotype of line Msm, “B” – a sensible genotype of line B6 and “H” – a heterozygous genotype.

A genotype information unit is the marker, which describes information of the chromosome fragment where the genotype is detected. A sample of genotype and phenotype values was processed by mathematical algorithms.

The purpose of the investigation was to find the marker or the group of markers responsible for sensitivity or resistance to TNF. The search is based on the hypothesis that a phenotype corresponds to a genotype.

At first sight, the purpose could be accomplished with the help of some statistical criteria (for example, Pearson’s phi-square). If only one group of genes on a chromosome (described by a marker) completely determined the sensitivity or resistance to TNF the problem would have a simple solution. However the previous investigations have not confirmed such assumption. Correlation between a genotype and sensitivity is a complex problem which cannot be solved by a single marker effect.

Another complication is that the problem statement does not agree with the “classical” one solved by the methods of correlation analysis, multiple regression analysis, ANOVA, pattern recognition [3, 4]. Typically approaches on single-marker analysis and multi-marker analysis are based on chi-squared test, likelihood-based test and logistic regression model [5, 6]. At that a considerable amount of research has focused on selection of appropriate computational methods and software packages for geneticists and other biomedical challenges [6, 7]. The graph approaches to present a human genome based on correlation and entropy are described in work [8]. Up to date statistical software [9] utilizes data mining algorithms with classification trees methods to provide hierarchic model of data. However, these algorithms are most suitable for the study of the predictor effect on a continuous result. In a case of large amount of predictors the methods create complex hierarchical structure which is hard to interpret and utilize.

Nevertheless, our work was focused on pattern building methods for modelling of the multiple discrete variables effect on a discrete result. These methods require the adaptation of standard methods and the development of particular methods and approaches.

This paper proposes several approaches to find markers combinations affecting sensitivity to TNF. They are based on

correlation analysis, theory of probability and graph theory. The developed algorithms were implemented and applied to the input data.

II. Problem Statement

The initial data were collected by the Laboratory of Molecular Genetics of Constitutive Immunity at Petrozavodsk State University. The sample consisted of genetic information contained in 44 markers of 20 chromosomes. The marker values were detected for 213 mice: 137 with a *sensitive* phenotype and 76 with a *resistant* phenotype.

Let's present the initial sample Ω into classes: Ω_R containing *resistant* objects and Ω_S containing *sensitive* ones. It is obvious that $\Omega_R \cap \Omega_S = \emptyset$ and $\Omega_R + \Omega_S = \Omega$.

Let's describe the genetic data for object with number j as a vector:

$$x_j = \left(x^1_{m_1}, x^1_{m_2}, \dots, x^1_{m_{n_1}}, x^2_{m_1}, x^2_{m_2}, \dots \right.$$
$$\left. \dots, x^2_{m_{n_2}}, \dots, x^{20}_{m_1}, x^{20}_{m_2}, \dots, x^{20}_{m_{n_{20}}}\right)_j, \quad (1)$$

where $j = 1,2,\dots,N$ and N– the sample volume, $x^h_{m_k}$ – genotype for marker on h-th chromosome with genetic distance equal m_k centimorgan, $x^h_{m_k} \in \{A; B; H\}$, $h = 1,2,\dots,20$, $k = 1,2,\dots,n_h$, n_h- the number of markers on h-th chromosome, $\sum_{h=1}^{20} n_h = n$.

Let $\boldsymbol{X}(H, M, X)$ denote a **pattern,** which is a combination of several values of markers:

$$\boldsymbol{X}(H, M, X) = \left(x^{h_1}_{m_1} = X_1, x^{h_2}_{m_2} = X_2, \dots, x^{h_n}_{m_n} = X_n\right), \quad (2)$$

where h_k – number of chromosome, $h_k \in H$, m_k – genetic distance of marker, $m_k \in M$, X_k – values of marker, $X_k \in X$, $k = 1,2,\dots,q$, value q – the length of pattern. If the object x has the combination of values (2), it means that «the object x corresponds to the pattern $\boldsymbol{X}(H, M, X)$».

As it was considered above, the problem is to find patterns with high connection with sensitivity or resistance to TNF.

III. Methodology

A. Hierarchic patterns

This method finds the patterns $\boldsymbol{X}(H, M, X)$ in accordance with some criteria of quality. The criteria can be: resistance probability maximization for the objects which correspond to the pattern:

$$P(R(x) = 1 | \boldsymbol{X}(H, M, X)) \rightarrow max, \quad (3)$$

– belonging of such probability to given interval:

$$P(R(x) = 1 | \boldsymbol{X}(H, M, X)) > P_{opt}, \quad (4)$$

where $P(\dots|\dots)$ – conditional probability, $R(x)$ – phenotype for object x (0 – sensitive, 1 – resistant), P_{opt} – the value of probability. We should use enough high values of probability to assure the high survive probability for the objects connected with the pattern [10].

The preliminary stage of the method was required to define the markers with high connection with sensitivity or resistance to TNF. The rank correlations [11] were used to evaluate the degree of association between markers and resistance to TNF.

Rank variables $x^h_m = \{-1; 0; 1\}$ were used to describe initial genetic data for marker m on chromosome h. The value "1" corresponds to genotype "A", "-1" corresponds to "B" and "0" to "H". Rank variable $r = \{0; 1\}$, where "1" corresponds to resistant phenotype, "0" to sensitive one. As it was mentioned above the basic hypothesis of the investigation was "phenotype corresponds to genotype". So, the significant positive correlations match to the case when the resistant genotype connects with the resistant phenotype [12].

Patterns with hierarchic structure were built by the following method. The markers having significant positive correlations were put in the "root" of the hierarchic patterns. In the process the values "A" were put in the pattern root for Ω_R class, value "B" for Ω_S class.

The markers with high frequency of combination with the "root" marker were linked to the "root". If the probability of combination was more than P_{opt} the combination became the "pattern" consisting of two markers. The second marker in the combination presented the second level of the hierarchic pattern. At the next step the markers which had not been used were added to the previous level markers. The iteration process of markers addition should be continued until the probability of pattern meets the condition (4). When the process was completed we had a set of hierarchic patterns consisting of $2,3,4,\dots,n$ levels. Such patterns had high association with the sensitivity or resistance to TNF.

B. Patterns as a graph

The method is based on frequency calculation of paired occurrence of marker values for each class of objects and on presentation the results as a graph [10].

At the first step we evaluated paired frequency p^α_{ij}, $\alpha \in \{S, R\}$ for each pair of marker values for each class of objects, $i, j = 1,2,\dots,3n$. The dimension $3n$ connected with three concerned genotypes. The results were presented by two $3n \times 3n$ matrices P_S and P_R for Ω_S and Ω_R classes respectively.

Then we calculated differences $\Delta^S_{ij} = p^S_{ij} - p^R_{ij}$ for Ω_S class and $\Delta^R_{ij} = p^R_{ij} - p^S_{ij}$ for Ω_R class, $i, j = 1,2,\dots,3n$.

The next step required selection of marker pair values (i, j) for each class with differences larger than the prescribed parameter $0 < \varepsilon < 1$. These pairs were presented as the edges of graphs, the paired frequency p^α_{ij}, $i, j = 1,2,\dots,3n$, $\alpha \in \{S, R\}$ were defined as the weights of the edges.

The next edges could be added to the graphs by selection of the markers k: such as $k = argmax_l\{\Delta^\alpha_{il}, \Delta^\alpha_{lj}\}$ for Ω_α, $\alpha \in \{S, R\}$. The graphs for each class were built separately with differences Δ^S_{ij} and Δ^R_{ij} for Ω_S and Ω_R classes respectively.

The process of selection and addition could be continued while the pairs with differences higher than ε were available. When the process was completed a set of patterns as a graph was obtained. The level of ε defined the degree of patterns association with sensitivity or resistance to TNF.

C. Results. Hierarchic patterns

At the preliminary stage of the method the markers with statistically significant positive correlations with resistance to TNF were defined. The results of calculation are summarized in Table 1.

The first column of the table contains marker descriptions by the form of "**D**<number of chromosome>**M**<genetic

distance>". The levels of correlation significance are indicated by P-level in the next column. The third column involves the values of Kendall τ and Spearman r_s rank correlations.

TABLE I
CORRELATIONS BETWEEN MARKERS AND RESISTANCE TO TNF

Markers	P-level	τ; r_s
D1M236	p=,00035	+0,22; +0,23
D1M132	p=,00233	+0,19; +0,2
D1M74	p=,07947	+0,23; +0,25
D8M6	p=,04497	+0,08; +0,08
D11M4	p=,05326	+0,12; +0,13
D11M99	p=,07845	+0,1; +0,1
D11M333	p=,01823	+0,1; +0,1
D15M224	p=,00384	-0,17; -0,18
D15M7	p=,07413	-0,08; -0,08
D1M236	p=,00035	+0,22; +0,23
D1M132	p=,00233	+0,19; +0,2
D1M74	p=,07947	+0,23; +0,25
D8M6	p=,04497	+0,08; +0,08
D11M4	p=,05326	+0,12; +0,13
D15M224	p=,00384	-0,17; -0,18
D15M7	p=,07413	-0,08; -0,08

Fig. 1 demonstrates that positive correlation matches to the case when the resistant genotype connects with the resistant phenotype. The values of correlations for marker «M1D132» are $\tau = +0.19$, $r_s = -0.2$.

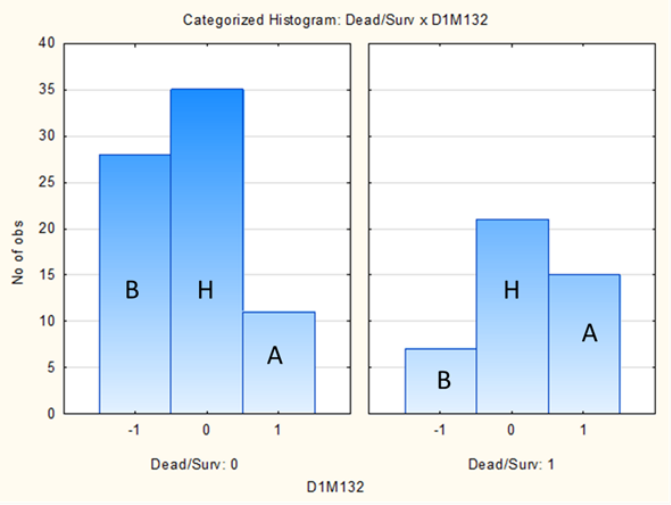


Fig. 1. Categorized histogram for marker «M1D132» by types of phenotype.

The frequency of markers values in Fig.1 show that the genotype «A» prevails in *resistant* phenotype group (right chart), and the genotype «B» prevails in *sensitive* phenotype group (left chart).

The opposite case is presented in Fig. 2. The values of correlations for marker «M15D224» are $\tau = -0.17$, $r_s = -0.18$.

The charts demonstrate that the genotype «B» prevails in the *resistant* phenotype group (right chart), and the genotype «A» slightly prevails in the *sensitive* phenotype group (left chart). This indicates that the negative values of correlation are in contrast to the hypothesis that the resistant genotype connects with the resistant phenotype.

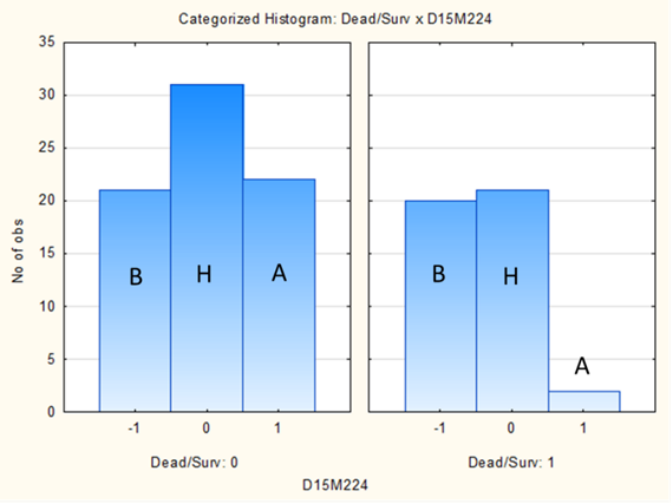


Fig. 2. Categorized histogram for marker M15D224 by types of phenotype.

The hierarchic patterns obtained by the algorithm were presented by the form illustrated in Fig.3.

Level 1		Level 2		Level 3		
		D10M7=H	0.680	D3M201=H	0.706	12
D1M236=B	0.338					
		D19M133=H	0.680	D3M110=H	0.824	14
				D3M211=H	0.706	12

Fig. 3. Hierarchic patterns for sensitive phenotype with value «B» of marker «D1M236» in the root.

The following hierarchic patterns are showed in Fig. 3:

(D1M236=B; D10M7=H; D3M201=H), 12 objects from Ω_S are corresponded to the pattern;

(D1M236=B; D19M133=H; D3M110=H) for 14 objects from Ω_S;

(D1M236=B; D19M133=H; D3M211=H) for 12 objects from Ω_S.

The hierarchic structure in Figure 3 also contains frequency of combinations:

the frequency of value D1M236=B is 0.338 among the objects Ω_S;

the frequency of value D10M7=H is 0.68 among the objects Ω_S with D1M236=B;

the frequency of value D3M201=H is 0.706 among the objects Ω_S with D1M236=B and D10M7=H.

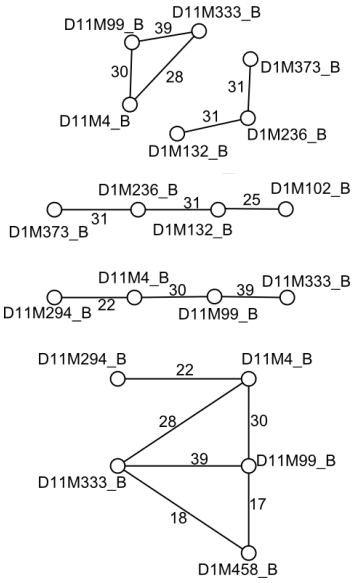
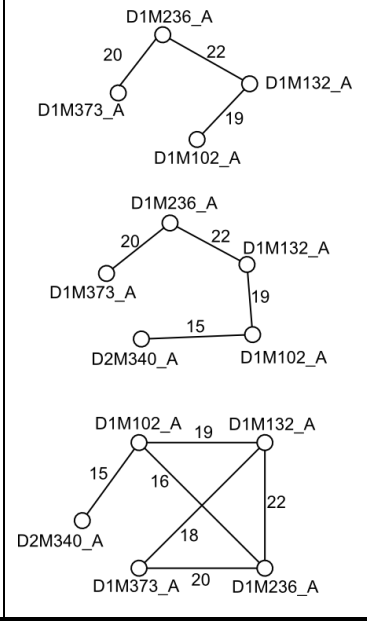
The fragment of 3-levels hierarchic patterns building results are presented in Table 2. We use P_{opt} equal 0.4 on 2[d] and 3[rd] levels.

Table 2 contains patterns with high number corresponding objects. The results show that the most significant patterns for sensitive phenotype include the markers with "B" genotype located on 1[st], 8[th] and 11[th] chromosomes. The patterns containing markers with negative values of correlation were also built. The patterns similar to (M15D224 = A; D15M7=A; D4M196=B) could confirm the contradiction to the hypothesis that the resistant genotype connects with the resistant phenotype for markers with negative values of correlation.

TABLE II
HIERARCHIC PATTERNS BUILDING RESULTS

Level 1		Level 2		Level 3		Number of obj.
Sensitive phenotype group						
D1M373=B	0.270	D1M132=B	0.676	D11M294=A	0.440	11
		D1M458=B	0.568	D8M13=B	0.524	11
D1M236=B	0.285	D1M132=B	0.795	D4M196=B	0.419	13
		D1M458=B	0.487	D6M274=A	0.526	10
D1M236=B	0.285	D1M458=B	0.487	D8M120=B	0.526	10
D1M132=B	0.307	D1M236=B	0.738	D4M196=B	0.419	13
		D1M373=B	0.595	D11M294=A	0.440	11
D1M102=B	0.270	D1M458=B	0.622	D8M6=B	0.435	10
		D1M236=B	0.486	D4M196=B	0.500	9
D1M458=B	0.350	D1M102=B	0.479	D8M6=B	0.435	10
		D1M373=B	0.438	D8M13=B	0.524	11
D11M294=B	0.299	D11M4=B	0.537	D10M71=B	0.455	10
		D11M333=B	0.415	D8M6=A	0.588	10
D11M4=B	0.336	D11M294=B	0.478	D10M71=B	0.455	10
		D8M120=B	0.413	D5M394=A	0.526	10
D11M333=B	0.307	D11M294=B	0.405	D8M6=A	0.588	10
D15M224=A	0.299	D15M7=A	0.780	D4M196=B	0.438	14
				D8M6=B	0.406	13
		D15M104=A	0.610	D4M196=B	0.440	11
Resistant phenotype group						
D1M373=A	0.329	D1M102=A	0.560	D15M224=B	0.500	7
		DXM234=A	0.520	D5M98=B	0.538	7
D1M236=A	0.368	D5N4=B	0.536	D16M189=B	0.400	6
		DXM234=A	0.536	D5M98=B	0.400	6
D1M132=A	0.395	D5M98=B	0.567	D7M246=B	0.412	7
				D8M120=B	0.412	7
D1M102=A	0.355	D2M340=A	0.556	D5M98=B	0.533	8
				D6M274=B	0.400	6
D1M458=A	0.303	D1M236=A	0.478	D15M224=B	0.545	6
		D2M340=A	0.478	D6M274=B	0.545	6
D11M294=A	0.289	D5M98=B	0.500	DXM249=A	0.545	6
		D11M4=A	0.500	D3M22=B	0.545	6
D11M4=A	0.289	D11M294=A	0.500	D3M22=B	0.545	6
D11M99=A	0.289	D9M155=A	0.545	D5M98=B	0.500	6
D11M333=A	0.316	D15M224=B	0.417	D1M373=A	0.600	6

TABLE III
PATTERNS BUILDING RESULTS

Combinations for Ω_S class	Combinations for Ω_R class.
	

The results also show that the most significant patterns for resistant phenotype primarily include the markers with "A" genotype located on 1st and 11th chromosomes. Furthermore the study of resistant phenotype patterns demonstrated the great variety of examined objects genotype.

D. Results. Patterns as a graph

As mentioned previously the method is based on frequency calculation of paired occurrence of marker values for each class of objects and on presentation the results as a graph. The part of graphs built on the base of pairs with highest differences Δ^S and Δ^R are presented in Table 3. The paired frequency p_{ij}^S, p_{ij}^R were used as the weights of the edges.

Combinations for Ω_S class demonstrate amalgamation of markers with genotype "B". Algorithm combines the markers of 1st and 11th chromosomes into a set of patterns, for example: (D11M99=B; D11M333=B; D11M4=B) or (D1M373=B; D1M236=B; D1M132=B).

The algorithm combines the markers with genotype "A" for Ω_R class, for example: (D1M373=A; D1M236=A; D1M132=A; D1M102 = A).

The patterns presented in Table 3 can confirm the accordance to the hypothesis that the resistant genotype connects with the resistant phenotype. Developed patterns demonstrate the occurrence of multiple correlations between markers within classes.

The revealed graphs also show that the most significant markers responsible for sensitivity or resistance to TNF are located on 1st and 11th chromosomes. Consequently the accordance of object with patterns for Δ^S and Δ^R classes should increase the level of sensitivity or resistance to TNF respectively.

IV. CONCLUSION

Our investigation was focused on development of methods which would allow us to determine the multiple variables effect on a discrete result. The developed methods were used to find the marker or a group of markers responsible for sensitivity or resistance to TNF. We used two types of patterns building methods in the study. The obtained results reveal some combination of markers which might significantly affect the genotype.

At the first stage the markers with significant positive correlations with resistance to TNF were defined. Such correlations were detected for 1st, 8th and 11th chromosome markers.

We found that the developed patterns have a great variety. Implementation of the hierarchic patterns building method showed that even patterns with high occurrence frequency matched to a few objects. The results provided evidence of significant correspondence of 1st and 11th chromosome markers with resistance to TNF. The same results were demonstrated by the patterns built according to the graph method.

The study of genetic information demonstrated the great variety of examined objects. The variety essentially complicated the process of finding a group of markers responsible for sensitivity or resistance to TNF.

Further work should therefore include a more detailed study of 1st and 11th chromosome markers and further development of pattern building methods.

ACKNOWLEDGEMENT

The authors would like to thank the Laboratory of Molecular Genetics of Constitutive Immunity at Petrozavodsk State University for collection and providing initial data. Also

we would like to express our gratitude to prof. A.N. Poltorak for ideas about methods and some useful recommendations.

REFERENCES

[1] B. Beutler, A. Poltorak "Sepsis and evolution of the innate immune response". Crit Care Med. 2001; 29(7 Suppl):S2-6; discussion S6-7.

[2] A. Kruglov, A. Tumanov, S. Grivennikov, Y. Shebzukhov, A. Kuchmiy, G. Efimov, M. Drutskaya, J. Scheller, D. Kuprash, S. Nedospasov. "Modalities of experimental TNF blockade in vivo: mouse models". Adv Exp Med Biol. 2011. 691:421-31.

[3] P. Armitage "Statistical Methods in Medical Research". Wiley. 2002.

[4] X.-H. Zhou, N. Obuchowski, D. McClish "Statistical Methods in Diagnostic Medicine". John Wiley & Sons. 2011.

[5] G. Zheng, Y. Yang, X. Zhu, R. C. Elston "Analysis of Genetic Association Studies". Springer, 2012.

[6] G. Clarke, C. Anderson, K. Zondervan "Basic statistical analysis in genetic case-control studies". Online: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3154648/.

[7] S. Lin, H. Zhao "Handbook on Analyzing Human Genetic Data: Computational Approaches and Software". Springer, 2009.

[8] E. Zeggini, A. Morris "Analysis of Complex Disease Association Studies: A Practical Guide". Academic Press, 2010.

[9] StatSoft Electronic Statistics Textbook. Online: http://www.statsoft.com/Textbook/Classification-Trees

[10] J. Han, M. Kamber, J. Pei "Data Mining: Concepts and Techniques". Elsevier. 2011.

[11] S.A. Aivazyn, I.S. Yenyukov, L.D. Meshalkin "Applied Statistic. Study of Relationships". Finansy i statistika. Moscow. 1985. (in Russian).

[12] S.A. Aivazyn, I.S. Yenyukov, L.D. Meshalkin "Bases of modelling and initial data processing". Finansy i statistika. Moscow. 1983.(in Russian).