Discriminative Power of Lymphoid Cell Features: Factor Analysis Approach[1]

Igor Gurevich[1], Dmitry Harazishvili[2], Irina Jernova[1],
Alexey Nefyodov[1], Anastasia Trykova[1], and Ivan Vorobjev[2]

[1] Scientific Council "Cybernetics" of the Russian Academy of Sciences
40, Vavilov str., 119991 Moscow, GSP-1, Russian Federation
igourevi@ccas.ru

[2] Hematological Scientific Center of the Russian Academy of Medical Sciences
4a, Novyi Zykovskii pr., 125167 Moscow, Russian Federation
ivorobjev@mail.ru

Abstract. The new results of the research in the field of automation of hematopoietic tumor diagnostics by analysis of the images of cytological specimens are presented. Factor analysis of numerical diagnostically important features used for the description of lymphoma cell nucleus was carried out in order to evaluate the significance of the features and to reduce the considered feature space. The following results were obtained: a) the proposed features were classified; b) the feature set composed of 47 elements was reduced to 8 informative factors; c) the extracted factors allowed to distinguish some groups of patients. This implies that received factors have substantial medical meaning. The results presented in the paper confirm the advisability of involving factor analysis in the automated system for morphological analysis of the cytological specimens in order to create a complex model of phenomenon investigated.

1 Introduction

In this paper, we describe new results of the research into automation of hematopoietic tumors diagnostics on the base of analysis of the images of cytological specimens. This work has been conducted since 2000 by the researchers of the Scientific Council "Cybernetics" of the Russian Academy of Sciences together with the researchers of the Hematological Scientific Center of the Russian Academy of Medical Sciences [1]. The necessary condition of such automation is the development of information technology for morphological analysis of the lymphoid cell nuclei of patients with hematopoietic tumors, which could be implemented in corresponding software system for automated diagnostics. The paper is devoted to the investigation of the numerical features used for the description of lymphoma cell nucleus by means of factor analysis. The method of factor analysis, which allows reducing and structur-

[1] This work was partly supported by the Russian Foundation for Basic Research (grant No. 01-07-90016, and 03-07-90406) and by Federal Target-Oriented Program "Research and Development in the Priority Directions of Science and Technology" in 2002-2006 (project No. 37.0011.11.0016).

A. Sanfeliu and J. Ruiz-Shulcloper (Eds.): CIARP 2003, LNCS 2905, pp. 298–305, 2003.

ing the initial data, proved to be efficient for the problem of morphological analysis of lymphocyte nucleus.

The paper is organized as follows. Section 2 contains a brief description of the developed information technology for the morphological analysis of the cytological specimens. Section 3 contains information about the initial data for factor analysis and about methods used for factor analysis. The results of factor analysis and some conclusions are presented in section 4. Note that the developed technology is described entirely in [4].

2 The Main Stages of the Morphological Analysis of the Blood Cells

The developed information technology for the morphological analysis of the cytological specimens includes the following stages of data preparation and analysis:

1. Creation of a database containing images of specimens of lymphatic tissues with isolated lymphocyte nuclei for patients with different lymphoid tumors.
2. Normalization of the images in order to compensate for different illumination conditions and different colors of stain used for the specimens.
3. Choice of features which capture morphological characteristics of lymphocytes nuclei useful for lymphoma diagnostics.
4. Calculation of values, statistical and qualitative analysis of the chosen features for the set of available nuclei.
5. Selection of features for generating feature descriptions of lymphoma cell nuclei.
6. Cluster analysis of the nuclei by using different subsets of the generated set of features.
7. Qualitative and quantative analysis of the obtained clusters.
8. Formation of a new feature space for description of the patients:
   - the ‘large’ clusters of the cell nuclei are selected;
   - the new features are relational numbers of patient’s nuclei that belong to the selected clusters.
9. Diagnosing the patients by the use of efficient recognition algorithms (for example, recognition algorithms based on estimate calculation [5]) applied to the feature descriptions developed in p.9.

2.1 General Characteristic of the Source Data

A base of photomicrographic images of lymphatic tissue imprints was created to select and describe diagnostically important features of lymphocyte nuclei images. The base contains 1585 photos of specimens of 36 patients. We choose 25 cases of aggressive lymphoid tumors (de novo large and mixed cell lymphomas (L) and transformed chronic lymphocytic leukemia (TCLL)) and 10 cases of indolent chronic lymphocytic leukemia (CLL). In one case the reactive lymphoid hyperplasia was diagnosed.

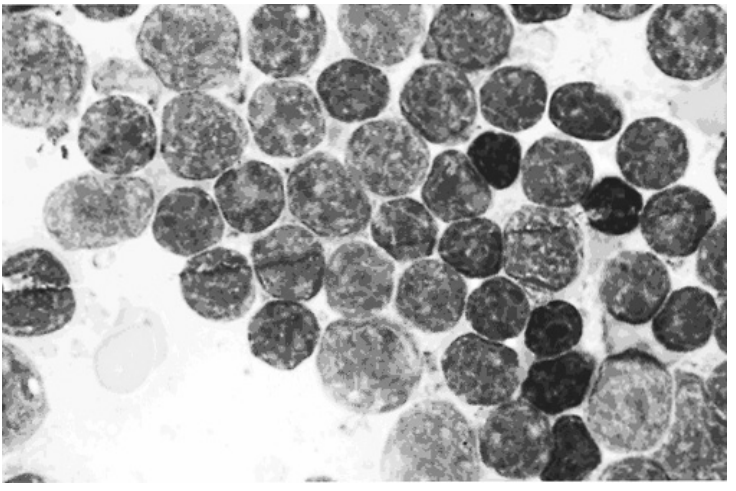

Fig. 1. Monochrome photomicrographic image of the footprint of lymphatic gland. Objective ×100.

The photos of specimens were collected and stored as the RGB-images in 24-bit TIFF format. Figure 1 shows monochrome picture of the slide of lymphatic gland. On the original RGB-images 4327 nuclei of lymphocytic cells important for diagnostics were indicated by experts. These nuclei were further segmented and analyzed.

2.2 Selection and Extraction of Features for Lymphocyte Nuclei Description

The principal property of the proposed technology is the generation of lymphocyte nuclei description by the features chosen and calculated from the images of specimens by the methods of image processing and analysis, and also by the methods of mathematical morphology and Fourier analysis.

During the morphological analysis of lymphocyte nuclei, hematologists use the following characteristics.

1. Nuclei size and density of different lymphoid cells in the specimen (presence of cells with nucleus larger than in most cells).
2. Nuclear form (round, oval, folded), presence of invaginations.
3. Textural features of chromatin (dispersed or condensed pattern; if dispersed – visual diameter of chromatin fibrils).
4. Presence or absence of nucleoli; if present – their number, size, form and location in nuclei (central or peripheral).

We provided formal equivalents of some of the above characteristics. They can be translated into geometrical and textural features of cell nuclei. Thus, the following 47 features were chosen to describe nuclei morphology:

1) an area of nucleus in pixels;
2) four statistical features calculated on nucleus brightness histogram (average, dispersion, 3rd and 4th central moments);

3) 16 granulometric features of nucleus;
4) 26 features calculated on the Fourier spectrum of nucleus.

2.3 Cluster Analysis of the Nuclei

To preprocess the images of the specimens and to calculate the nuclei' features, a program was developed which uses the libraries of the "Black Square" software system [3]. After calculation of the features, their statistical and qualitative analysis was carried out. Conducted analysis allowed us to select several groups of features, which further yielded diagnostically-perspective taxonomies.

The main results of the lymphocytic cells investigations are based on their cluster analysis using developed feature space and its subspaces with the help of FOREL algorithm [2].

A number of sets of clusters were received by using different sets of features and different values of parameter of FOREL algorithm. Obtained sets of clusters were evaluated using various criteria (e.g., the number and the size of clusters in each set, the character of nuclei distribution in large clusters, etc.). Several "interesting" sets of clusters were selected which have good formal characteristics of their clusters and are promising for interpretation by hematologists.

The obtained results showed that:

1) the set of diagnostically important nuclei of the patients with considered lymphoid tumors is substantially heterogeneous since different clusters in it were clearly identified;
2) clusterization of the lymphocytic nuclei using developed feature set is important from the medical standpoint;
3) formal nucleus characterization by the developed feature set corresponds well with its qualitative morphological description and serves as a basis for development the automated software systems of morphological analysis of the blood cells and diagnosis of hemoblastoses;
4) the suggested technology provides transition from the diagnostic analysis of lymphocytes nuclei to diagnosing patients with hematopoietic tumors by the use of pattern recognition techniques.

3 Factor Analysis of the Features of Lymphocyte Nuclei

As it is known, factor analysis allows one to estimate the dimensionality of a set of observed variables and to determine the structure of interconnections between these variables. Factor analysis can be used to replace a large set of observed variables with a smaller set of new hypothetical variables called factors. These factors are treated as principal variables that truly describe investigated phenomenon.

The initial data were represented as a table with the rows corresponding to the nuclei' feature descriptions. Factor analysis was conducted on the whole of given nucleus, on different combinations of nucleus taken from patients with different forms

of hemoblastoses (L, CLL and TCLL), and on 5 large clusters, extracted in the set of nucleus by means of taxonomy method.

The tables of nucleus feature descriptions were normalized by columns by the rule

$$z_{ij} = \frac{y_{ij} - \overline{y_j}}{s_j}, \tag{1}$$

where z_{ij}, y_{ij} are the normalized and primary values of the *j*th feature of the *i*th nucleus, s_j is a standard deviation of the *j*th feature, $\overline{y}_j$ is an average value of the *j*th feature. As a result of such transformation the features' variances became unit. Then, the correlations of the features were calculated in each group of nuclei. Factor analysis was applied to the corresponding reduced correlation matrices with the communities on the principal diagonal. In some cases iterative procedure was applied for evaluation of communities, or squared coefficient of multiple correlation of features with the other ones was considered. The extraction of the factors was realized by means of 3 methods, namely by means of principal-factor method, centroid method and maximum-likelihood method. The quantity of the factors was determined as a result of combining the Kaizer criterion and the scree-test. The varimax-rotate strategy was applied with the purpose of obtaining the contansive interpretation of factors.

It is known that factor problem has an ambiguous solution depending on restrictions imposed on (the choice of factorial method). But considering that a single simple factor explanation lay in the very data, different methods should yield nearly the same mappings, and it signifies that it is possible to establish a certain correspondence between factors from different solutions. Thus, next stage of analysis was the search of correspondences between factor mappings received with the help of 3 different methods. Then average factor loadings of each feature under 3 methods were calculated, and factors were ranked by a maximal percent of variance explained by them.

The conducted factor analysis allowed classifying different numerical features according to the cross correlation in independent groups, defining dominant factors. As a matter of fact each new factor proved to be a linear combination of several initial features, signed with high loadings (exceeding 0,7) for this factor (the most informative features). The additional analysis revealed that a choice of the value 0,7 as a threshold was advisable as long as the combination of features with high loadings did not vary significantly under decrease of this value.

4 The Results of Factor Analysis

The results of factor analysis are presented in Table 1. The factor mappings received on the different groups of cell nucleus are presented in the columns of Table 1, and the factors of the same significance level are presented in the rows. The increase of row number corresponds to a reduction of statistical significance of the factor, which is determined in turn by decrease of the variance of this factor. Thus, we come to the following conclusions:

1. The initial set of features (47 elements) breaks at the average into 8 significant groups – factors. There are some cases where 3 factors of the greatest information

density (which explain the largest part of the total variance of the features) combine into a single factor (general distribution of features by factors) in the group of patients with TCLL diagnosis and in the first cluster; and into 2 factors as well – in the groups of patients with CLL and L diagnosis. Due to mentioned above the quantity of important factors varies. The increase of the quantity of factors on the samples of nucleus of the smaller size may be explained by the following rule: if the dimensionality of the considered sample decreases until a certain moment the accuracy of its description with the help of the greater number of factors increases.

2. The factor mappings explain at the average 75% of the total variance of the features.
3. The analysis conducted allows establishing the following classification of the features.
   a. The main factors include nearly all texture features, the feature with the number 1 (an area of nucleus in pixels) and 3 granulometric features with the numbers 14, 15, 16 (the general number of light grains in nuclei, the number of grains with typical size and with minimal size respectively). The features included in the main factors are of the greatest importance for hematopoietic tumors diagnostics and these very features should be considered in the first place.
   b. The second dominant factor includes the statistical features with the numbers 3, 4, 5 – variance, 3rd and 4th central moments calculated on nucleus brightness histogram respectively.
   c. The features with the numbers 6, 7 (average and variance calculated on nucleus size histogram respectively) fall into the same factor with a minor statistical significance as a rule.
   d. Granulometric features with the numbers 10, 18 (the number of grains with sizes corresponding to local maxima and local minima of the constructed functions) occur together as well.

   The coefficients of cross correlation of the features are high enough in each factor.
4. It appeared that there are features that are inessential (have minor factor loadings) in all kinds of analysis, so these features bring in a little of new information. All these variables belong to the class of granulometric features of the nuclei and have the numbers 8, 9 - 3rd and 4th central moments calculated on nucleus brightness histogram, 11, 12, 13 – typical, minimal and maximal size of light grains in the nuclei respectively. The features with the numbers 9, 11, 12, and 13 have minor pair correlations with the other parameters stably (<0,7).
5. The clusters and groups of patients are distinguished by the factors. It is important to note that some groups of nucleus contain features which don't occur in the set of the other groups.

	Groups of patients								Cluster ID									
	All patients		CLL diagnosis		TCLL diagnosis		L diagnosis		1		3		4		11		13	
1	1	0,8684	2	0,9405	22	0,95088	22	0,918	1	0,807	30	0,879	2	0,871	30	0,779	2	0,7973
	14	0,8364	31	0,9184	23	0,94704	23	0,842	15	0,774	33	0,923	30	0,871	33	0,935	4	-0,768
	15	0,8665	33	0,7420	24	0,86301	26	0,91	16	0,778	38	0,917	31	0,839	38	0,838	22	-0,921
	16	0,8624	39	0,7405	25	0,70666	27	0,823	22	0,965	39	0,91	32	0,839	39	0,931	23	0,8467
	20	0,8102	41	0,7415	26	0,95364	29	0,903	23	0,96	41	0,923	33	0,836	41	0,933	24	0,9189
	22	0,8010	43	0,8112	27	0,87859	34	0,912	24	0,951			38	0,878			25	0,9561
	23	0,8263	44	0,8922	28	-0,7286	35	0,702	25	0,96			39	0,843			31	0,7416
	24	0,8169	46	0,8086	29	0,95396	37	0,903	26	0,942			41	0,836			43	0,7505
	25	0,7102	47	0,8884	30	0,91999	42	0,875	29	0,961			43	0,802			44	0,7263
	26	0,7453			31	0,8554	45	0,875	30	0,962			44	0,822			46	0,7475
	29	0,8183			32	0,90004			31	0,908			46	0,801			47	0,7083
	30	0,8299			33	0,88044			32	0,908			47	0,828				
	31	0,8536			34	0,95437			33	0,921								
	32	0,8266			35	0,86902			34	0,942								
	33	0,8771			37	0,95392			35	0,874								
	34	0,7439			38	0,9215			37	0,961								
	35	0,7808			39	0,88239			38	0,962								
	37	0,8184			40	0,75237			39	0,929								
	38	0,8370			41	0,87916			40	0,861								
	39	0,8848			42	0,93716			41	0,923								
	40	0,8188			43	0,82977			42	0,937								
	41	0,8782			44	0,77761			43	0,941								
	42	0,8088			45	0,93711			44	0,88								
	43	0,8864			46	0,82985			45	0,937								
	44	0,8846			47	0,78202			46	0,94								
	45	0,8086							47	0,886								
	46	0,8860																
	47	0,8793																
2	27	0,6304	22	0,82808	1	0,71857	2	0,749	2	0,535	22	-0,944	22	-0,762	22	-0,928	33	0,9615
			23	0,96263	14	0,70185	31	0,799			24	0,943	24	0,762	24	0,922	39	0,9635
			24	0,84334	15	0,7106	32	0,73			25	0,936			25	0,899	41	0,9585
			25	0,96846			33	0,714										
			29	0,96126			41	0,714										
			30	0,74572			43	0,711										
			37	0,95388			44	0,811										
			38	0,7392			46	0,71										
							47	0,816										
3	3	-0,7199	3	0,80844	3	0,7074	1	0,777	3	-0,72	1	0,828	26	-0,837	26	-0,902	28	-0,943
	4	-0,6700	4	0,77032			14	0,703	4	-0,736	15	0,798	34	-0,833	34	-0,902	32	0,7686
	5	-0,8217	5	0,88762					5	-0,842	16	0,794					36	-0,945
			15	0,70442							26	-0,903					40	0,7416
											34	-0,903						
4	2	0,6069	27	-0,8944	2	0,77808	6	0,897	6	0,71	3	-0,864	3	0,832	15	0,839	15	0,8988
	6	0,8133	28	0,88313					7	0,838	5	-0,849	4	0,781	16	0,841	16	0,8955
			32	-0,7118									5	0,88				
5	10	-0,8827	10	0,91554	6	-0,8389	3	0,888										
	18	-0,7255	18	0,86401			5	0,914										
6			6	0,84595	10	0,90486	10	-0,887			6	0,829	6	0,786	6	0,697	26	0,9271
			7	0,69465	18	0,82595	18	-0,801					7	0,864	7	0,812	34	0,9147
7			26	0,78856			25	0,716			42	0,947	42	-0,888	3	0,884	7	0,8834
			34	0,78194							45	0,948	45	-0,885	5	0,825		
8			12	0,55118			19	-0,688							42	-0,886	42	-0,968
							21	-0,667							45	-0,885	45	-0,972
9			42	0,87081											29	-0,637	29	0,4215
			45	0,87081											37	0,633	37	-0,403
10																	3	0,7141

Table 1. The results of factor analysis of the features for different groups of the nuclei. First column contains factor numbers, in the rest of columns first number is the feature number while the second number is the value of the respective factor loading.

5 Conclusion

The method of factor analysis, which allows reducing and structuring the initial data, proved to be efficient conformably to the problem of morphological analysis of lymphocyte nucleus. It confirmed that the proposed feature space, reflecting morphological characteristics of lymphocyte nucleus used in diagnostics, has a sufficiently simple factor structure. The main goals of factor analysis were achieved, since we succeeded in reducing the feature set composed of 47 elements at least to 8 informative factors and in making a classification of the features proposed. The important result is that the extracted factors allow to distinguish some groups of patients. This implies that received factors have contansive medical meaning. The results presented above are the prerequisites for involving factor analysis in the automated system for morphological analysis of the cytological specimens in order to create a complex model of phenomenon investigated.

In future we intend to carry out the factor analysis on the other samples of patients in supplemented feature space for the purpose of conformation of existence of extracted factors, and to exploit new methods of factor analysis as well.

References

1. Churakova, J.V., Gurevich, I.B., Hilkov, A.V., Jernova, I.A., Kharazishvili, D.V., Nefyodov, A.V., Sheval, E.V.: Selection of Diagnostically Valuable Features for Morphological Analysis of Blood Cells. Pattern Recognition and Image Analysis: Advances in Mathematical Theory and Applications. **2** (2003) 382 –383
2. Elkina, V.N., Zagoruiko, N.G.: Some Classification Algorithms Developed at Novosibirsk. In: Simon, J.C. (ed.): Intelligence Artificielle, Reconnaissance des Formes. R.A.I.R.O. Informatique/Computer Science. **1** (1978) 37–46
3. Gurevich, I.B., Khilkov, A.V., Murashov, D.M., et al.: Black Square Version 1.0: Programm Development System for Automation of Scientific Research and Education. Pattern Recognition and Image Analysis: Advances in Mathematical Theory and Applications. **4** (1999) 609–634
4. Gurevich, I.B., Harazishvili, D.V., Jernova, I.A., Nefyodov, A.V., Vorobjev, I.A.: Information Technology for the Morphological Analysis of the Lymphoid Cell Nuclei. Proceedings of the 13th Scandinavian Conference on Image Analysis (SCIA 2003), Sweden, June 29 - July 2, Springer, (2003) 541-548
5. Zhuravlev, Yu.I., Gurevitch, I.B.: Pattern Recognition and Image Recognition. Pattern Recognition and Image Analysis: Advances in Mathematical Theory and Applications in the USSR. **2** (1991) 149–181