CHAPTER 52

# Prediction and Precise Diagnosis of Diseases by Data Pattern Analysis in Multiparameter Flow Cytometry: Melanoma, Juvenile Asthma, and Human Immunodeficiency Virus Infection

**Günter Valet,[*] Hanna Kahle,[*] Friedrich Otto,[†] Edeltraut Bräutigam,[‡] and Luc Kestens[§]**

[*] Cell Biochemistry Group
Max-Planck-Institut für Biochemie
D-82152 Martinsried, Germany

[†] Fachklinik Hornheide
Abteilung für Tumorforschung
D-48157 Münster, Germany

[‡] Pathologisches Institut
Klinikum Görlitz
D-02828 Görlitz, Germany

[§] Prince Leopold Institute for Tropical Medicine, Pathology & Immunology
B-2000 Antwerp, Belgium

I. Introduction
II. Material and Methods
    A. Melanoma
    B. Asthma
    C. HIV Infection
    D. Flow Cytometric List Mode Analysis
    E. Data Pattern Classification
III. Results
    A. Melanoma

METHODS IN CELL BIOLOGY, VOL. 64

0091-679X/01 $35.00

B. Asthma
C. HIV Infection
IV. Discussion
References

## I. Introduction

The frequent use of multiparameter flow cytometry in the clinical or research laboratory together with the determination of a significant number of humoral biochemical parameters from blood serum, urine, or spinal fluid of patients generates a very substantial amount of information. Such information is only selectively evaluated at present, for example, according to the frequency of lymphocyte subpopulations, cell lineage assignments, abnormal immunophenotype, or cell activation marker description based on flow cytometric histograms or flow cytometric list mode data, as well as according to humoral or clinical parameters. Owing to the lack of suitable tools, the real potential of the multidimensionality of the measured information remains largely inaccessible.

This situation is highly unsatisfactory, considering the theoretical potential of cytometric information, for example, for the prediction of further disease course in individual patients. Cytometrically determined biochemical parameter patterns from directly or indirectly affected cellular systems or organs are of substantial interest for these purposes, because they are typically collected at the very spot of disease action and should therefore represent prime information carriers for disease course predictions, given the generation of diseases from biochemical deviations in cellular systems or organs. Predictions are preferable to statistical disease prognosis estimation. Although statistical disease prognosis is sufficient for therapy development and monitoring, it is of little value for the individual patient as well as for individualized therapy schemes.

The elaboration of general principles for disease course predictions in the clinical environment constitutes an important challenge for optimization of a patient's disease management. There are multiple methodological choices for the determination of structural or functional cell biochemical parameters and also for data evaluation by mathematical result modeling or by algorithmic principles. The task consists therefore of the rational selection of optimal biomolecular parameter patterns and result evaluation strategies for predictive medicine.

The high amount of data from flow cytometric measurements prompted the earlier development of the CLASSIF1 algorithm (Valet *et al.*, 1993) to assure the fast, exhaustive, and unbiased extraction of discriminant biomolecular data pattern from any kind of flow cytometric or other multiparameter data on typical personal computers, that is, out of hundreds or thousands of data columns. CLASSIF1 data classifications require specific and precise measurements accord-

ing to standardized methods, but no mathematical preconditions or assumptions have to be fulfilled. The resulting classifiers are robust and interlaboratory portable. The intellectual analysis of the selected discriminant data patterns of the classifiers favors scientific hypothesis formation through intuitive result presentation. Thus, large scale information extraction from multiparameter data is available, as an important precondition for individualized disease course predictions in patients.

Data analysis may simultaneously concern flow cytometric, humoral biochemistry, or clinical patient data in order to determine the most discriminant parameter pattern from the totality of the available information. In most instances, only a comparatively small amount of the entire information, that is, typically between 0.5 and 20%, assures the required discrimination.

The use of neural network (Frankel *et al.,* 1996, 1989; Boddy *et al.,* 1994; Molnar *et al.,* 1993; Ravdin *et al.,* 1993), principal component (Leary, 1994), cluster (Verwer and Terstappen, 1993; Demers *et al.,* 1992; Terstappen *et al.,* 1990; Schut *et al.,* 1993), or discriminant and statistical (Davey *et al.,* 1999; Hokanson *et al.,* 1999; Molnar *et al.,* 1993; Rothe *et al.,* 1990) analysis, hierarchical classifiers (Decaestecker *et al.,* 1996), classification and regression trees (CART, Beckman *et al.,* 1995), as well as knowledge based systems (Thews *et al.,* 1996; Diamond *et al.,* 1994) or fuzzy logic (Molnar *et al.,* 1993) describe major other approaches to multiparameter data analysis in cytometry and in the clinical laboratory. Difficulties of handling high parameter numbers, a need for mathematical assumptions, nonintuitive result presentation, problems with missing values, as well as complexity of implementation and operation have so far not led to a widespread application of these methodologies in the clinical or biomedical research environment.

It is the intention of this study to show the potential of the algorithmic CLASSIF1 approach for multiparameter data analysis in various clinical conditions.

## II. Material and Methods

### A. Melanoma

Clinical parameters, such as tumor thickness (TD, mm), tumor invasion depth into skin layers (LE, Clark level), TK as the mean of TD + LE, tumor ulceration (UL, 1 = no, 2 = yes) as well as flow cytometric DNA ploidy (euploid = 1, aneuploid = 2), and percentage of S-phase cells, that is, a total of six parameters were determined from surgery material and were available for 499 melanoma patients who either had survived 10 years after tumor surgery (A) or died (B) within this time period. The melanomas were localized on different parts of the dermal integument. For flow cytometry, 20–100 mg of tumor tissue was minced with a razor blade, incubated in a 2.1% citric acid, 0.5% Tween 20 solution for 20 min at 22°C, and centrifuged; the pellet was fixed with 70% ethanol, and

resuspended in citric acid/Tween 20 solution. The cell nuclei preparation was stained for 30 min in the presence of 1.75 $\mu$g/ml DAPI (4′,6-diamidino-2-phenylindole) in a 7.1% $Na_2HPO_4$ solution (Partec, Münster, Germany) (Otto *et al.*, 1981). The DNA fluorescence of the cell nuclei was measured with a PASII flow cytometer (Partec) using a HBO-100 high pressure mercury arc lamp with a UG1 (Schott, Mainz, Germany) fluorescence excitation filter and a GG435 (Schott) fluorescence emission filter for the determination of DAPI fluorescence. The coefficients of variation (CV) of the DNA distributions were between 1.5 and 3.1%, which is essential for the sensitive detection of DNA aneuploidy. The S-phase fraction was calculated from the DNA distribution according to a rectangular S-phase model.

### B. Asthma

Data from 40 juvenile asthma patients with a mean age of 10.00 ± 0.63 years (2.4–16.5 years) and 18 healthy reference children with a mean age of 10.65 ± 0.67 years (4.6–16.4 years) were processed. Available were 49 clinical chemistry parameters per patient as well as 103 lymphocyte frequency and relative fluorescence intensity values obtained by quadrant statistics from two-color whole blood lyse–nonwash direct mouse monoclonal antibody immunofluorescence assays using the following antibody combinations: CD4/8, CD3/19, CD3/HLA-DR, CD3/16+56, CD25/3, CD71/3, CD4/45RA, CD45RO/4, CD62L/4, CD4/29, CD57/8, CD8/11b, CD5/19, CD21/19, CD62L/20, IgG1/IgG2. Furthermore, list mode files from CD45/14, CD4/29, CD4/8, CD56/8, CD3/56, CD25/3, CD3/HLA-DR, CD71/3, CD3/19, CD5/19, and IgG1/IgG2 assays, prepared as outlined, were collected for 10,000 nucleated cells in flow cytometry standard (FCS) 1.0 format. Measured were fluorescein isothiocyanate (FITC) and PE (phycoerythrin) immunofluorescence as well as forward (FSC) and perpendicular (SSC) scatter light with a Becton Dickinson FACScan (Becton Dickinson, Heidelberg, Germany) flow cytometer.

### C. HIV Infection

Data of two-parameter peripheral blood lyse–nonwash FITC/PE immunophenotype measurements (CD45/14, CD3/16+56, CD2/19, CD45RA/4, HLA-DR/CD8, CD8/38, CD26/8, and CD26/4) of seronegative or human immunodeficiency virus (HIV)-infected seropositive patients were available as BD-FACScan (Becton Dickinson, Erembodegem, Belgium) list mode files as well as in the form of a 23-parameter dBase3 data base containing percent cell frequency values, manually extracted from two-parameter FITC/PE histograms, gated for lymphocytes by FSC/SSC.

### D. Flow Cytometric List Mode Analysis

The list mode files were analyzed with the CLASSIF1 list mode analysis software (Valet *et al.*, 1993). In short, two-parameter FITC/PE histograms of

FSC/SSC-gated lympho-, mono-, and granulocytes were evaluated by quadrant analysis for percent cell frequency, FITC and PE fluorescence intensity, fluorescence ratio, and relative FITC and PE antibody (Ab) surface density (fluorescence/square root of FSC) of the various cell populations using fixed fluorescence thresholds at one-third of the four decade logarithmic scale (channel 85 on a 256 channel scale). The FSC/SSC gates, in contrast, were autoadaptive for the three cell populations such as to always comprise more than 95% of the nucleated cells by nonoverlapping polygons. The calculated parameters of the FSC/SSC, FSC/FITC, and FITC/PE histograms as well as of the four quadrants of the FITC/PE histograms were data based (Valet and Höffkes, 1997) such that 34 data columns per leukocyte population, that is, a total of $3 \times 34 = 102$ data columns were available for lympho-, mono-, and granulocytes per measurement instead of only 12 parameters in case of cell frequency evaluation. To remain comparable with the cell frequency quadrant analysis, only data of single or double fluorescence positive quadrants were further classified; in other words, parameters of the fluorescence double negative cell population were not evaluated. This reduced the available parameters in each two-parameter immunophenotype for lympho-, mono-, and granulocytes to $3 \times 23 = 69$ columns for the subsequent data pattern classification.

## E. Data Pattern Classification

The results of the three studies were classified with the CLASSIF1 data pattern analysis algorithm (Valet and Höffkes, 1997; Valet *et al.*, 1993). The data were *a priori* assigned to either the learning set or to the embedded test set, such that patients 1, 5, 10, 15, . . . , of each classification category remained unknown to the algorithm during the learning phase. The algorithm proceeds as follows: Paired percentiles, for example, 10 and 90% percentiles for the value distributions of the learning set reference samples in each data base column are determined. Subsequently, all values of each data base column, that is, the values of the reference as well as of the abnormal samples, are transformed into triple matrix characters by assigning: "−" to values below the lower percentile, "0" to values between the percentiles, and "+" to values above the upper percentile. The resulting triple matrix replica of the numeric data base serves for the subsequent iterative data classification.

The algorithm optimizes during the learning phase the sum of the diagonal values of the confusion matrix, established between, for example, the known disease categories of patients on the ordinate and the computer determined classification of these disease categories on the abscissa. Ideally, that is, when all samples are correctly classified by the algorithm, the values in each of the diagonal boxes of the confusion matrix are 100%, while 0% values are encountered in the nondiagonal boxes. Since this ideal condition is not present at the beginning of the iterative optimization, the algorithm sequentially excludes either single data base columns or paired combinations of columns with any other of the data base columns temporarily from the classification process. Once all permutations

have been processed and no further improvement is reached, data base columns that had improved the classification result when excluded either alone or in combination with another column are permanently disregarded for the classification process; thus, only discriminant data base columns survive the selection process. The most frequent triple matrix character for each of the selected data columns of the learning set is entered into the classifier mask for each classification category. The classifier mask of the reference samples contains typically only 0 values because 0 values are the most frequent triple matrix characters in the value distributions of the reference samples from which the percentiles were calculated; for example, for the 10/90% percentile pair, 80% of triple matrix characters are 0, 10% are − and 10% are +.

The best possible classification is determined by successively classifying a data set separately for the different percentile pairs 10/90%, 15/85%, 20/80%, 25/75%, and 30/70% as well as cumulatively such that the most discriminant percentile pair for each data base column is used for the finally learned classifier. The three data sets on melanoma, juvenile asthma, and HIV infection were integrally processed, that is, without exclusion of samples.

As a check for a learned classifier, each patient of the learning set is reclassified according to the highest positional coincidence of the patient classification mask with anyone of the classifier masks at the finally selected classification conditions. Equal coincidence frequency for two classifier masks results in a double classification. Double classifications may represent either a biological transition state or a classification error, for example, in the case of small learning sets. The reclassification of the learning set samples permits a quick visual check of the sample triple matrices for systematic deviations (e.g., with time or on change of reagents) from the classifier masks.

The quality of a learned classifier is judged in a standardized way by the average recognition index (ARI) and by the average multiplicity index (AMI). The ARI is calculated as the sum of the diagonal values of the confusion matrix divided by the number of classification categories. It should be higher than 80% for clinical purposes. The AMI is a measure for the average frequency of assignment of more than one classification category to a sample. The AMI is ideally 1.00 in the absence of multiple classifications, and it is 1.1, 1.2, or 1.33 in case every tenth, fifth, or third example on average is assigned a double classification. AMIs between 1.0 and 1.2 are acceptable in practice. The AMI is calculated as the sum of the classification values in all lines of the confusion matrix divided by the number of classification categories, followed by a further division by 100. All AMIs of the subsequent classification were always 1.00, that is, no multiple classifications occurred.

The classification coincidence factor (CCF) indicates the coincidence of the individual patient/sample classification triple matrix with the best fitting classifier mask. The CCF is used to identify "unknown" sample classification masks that have a lower CCF than the lowest CCF observed during the learning process. Low CCFs may occur through inclusion of wrong samples, systematic errors during parameter

measurement, or because of missing values. Although systematically deviating test set samples are definitively rejected, samples with missing values are manually classified according to the best positional coincidence with any one of the classifier masks (e.g., patient 0001 in Fig. 2). No sample in all the processed samples of this study had to be excluded because of systematic deviations.

## III. Results

### A. Melanoma

The 499-patient data set consisted in sequence of 135 male and 216 female 10-year survivor (A) and of 83 male and 65 female nonsurvivor (B) patients. The first 75 male and 76 female (A) as well as 76 male and 64 female (B) patients were selected as the learning set with the first unknown test set. The learning set of 231 patients [59 male/61 female (A), 60 male/51 female (B)] contained the first unknown test set of 60 patients [16 male/15 female (A), 16 male/13 female (B)] in embedded form as outlined in Section II,E. The remaining 208 patients [60 male/140 female (A) together with 7 male/1 female (B)] served as the second unknown test set for the learned classifier. The total test set contained 268 patients.

The single-parameter sensitivity for correct nonsurvivor prediction was checked prior to data pattern analysis as a reference for data classification improvement by data pattern analysis. Single parameter sensitivity at 90% specificity for the identification of survivors was 53.7, 31.3, 50.9, 38.0, 38.1, and 21.7% for parameters TD, LE, TK, UL, AN, and SP. Individually, the values were too low for clinical predictions.

A first data pattern classification aiming at the discrimination of melanomas according to their location on the dermis was not successful (results not shown). In a second attempt it was investigated whether sex difference for survival existed in the provided data set. Due to identity of the classifier masks for male and female patients at the percentiles 10–90%, 15–85%, 20–80%, 25–75% and 30–70%, no distinction for the available parameters exists between male and female patients with regard to survival. Data from male and female patients can therefore be classified together in the search for melanoma-dependent differences in post tumor surgery survival (Table I). The impossibility of distinction between male and female patients is indirectly reflected by closely similar means and SEMs for each one of the investigated six parameters (Table II). Concerning the distinction between 10-year survivor and nonsurvivor patients, the optimal classifier (Table I) provides predictive values of 80.3% for survivors and 79.8% for nonsurvivors. The selected classification parameters are increased TD, TK, and SP. Only half of the provided parameters are selected for classification, although all six parameters are significantly increased for nonsurvivors (Table II). The triple matrix patterns of the three selected parameters for survivor (A) and nonsurvivor (B) (Fig. 1) show that the classification result remains in many instances correct,

**Table I**
**Melanoma: 10-Year Survivors and Nonsurvivors**

| Clinical outcome | Number of patients (*n*) | CLASSIF1 prediction (%)[a] | | |
|---|---|---|---|---|
| | | Survivor | Nonsurvivor | Specificity/ sensitivity |
| A. Learning set | | | | |
| Survivor | 120 | **81.7** | 18.3 | 81.7 |
| Nonsurvivor | 111 | 21.6 | **78.9** | 78.9 |
| Negative/positive predictive values | | 80.3 | 79.8 | ARI 80.0 |
| B. First unknown test set | | | | |
| Survivor | 31 | **80.6** | 19.4 | 80.6 |
| Nonsurvivor | 29 | 20.7 | **79.3** | 79.3 |
| Negative/positive predictive values | | 80.6 | 79.3 | ARI 80.0 |
| C. Second unknown test set | | | | |
| Survivor | 200 | **75.5** | 24.5 | 75.5 |
| Nonsurvivor | 8 | 12.5 | **87.5** | 87.5 |
| Negative/positive predictive values | | 99.3 | 85.4 | ARI 81.5 |

[a] 20–>30% optimized percentile thresholds (data base MELA6.BI4/.BI6).

although not all three parameters coincide with the classifier mask for either survivor or nonsurvivor. In other words, the individual patient classification is robust against a certain degree of noncoincidence with the best fitting classifier mask. The CCF is 0.67, that is, two out of three parameters have to match with the best fitting classifier mask for technically valid classifications.

The embedded first test set is classified with predictive values of 80.6% for the 31 survivors and with 79.3% for the 29 nonsurvivors (Table I). The second test set is classified with predictive values of 99.3% for the 200 survivors and with 85.4% for the 8 nonsurvivors (Table I). The classification of the two unknown test sets shows that the CLASSIF1 classifier provides a robust classification of unknown samples as an important quality criterion for multiparameter data classifiers.

## B. Asthma

Data pattern analysis of the clinical chemistry parameters provides a significantly better discriminatory result (Table III) than single parameter discrimination (Table IV). A sensitivity of 70% for the recognition of asthmatic children at 100% specificity for the identification of the healthy reference children is obtained. The respective positive and negative predictive values, that is, the correct prediction of the asthmatic and healthy children from the CLASSIF1 determined optimal parameter pattern (Table III), are 100 and 60% for the learning set. Seven of the 49 clinical chemistry measurements [i.e., thrombocyte (TRCS) and eosinophil (EOS) counts, aspartate aminotransferase (ASAT), thyroid-stimulating hormone (TSH), ferritin, IgE, and $\beta$-globulin] were selected by the CLASSIF1 algorithm. The data pattern classification of the clinical chemistry parameters for the test set patients

**Table II**
**Melanoma Parameters**[a]

| Patients | Number ($n$) | Diameter (TD, mm) | Infiltr.(LE) (arb.units) | TK = (TD + LE)/ 2 (mm) | Ulceration (UL) | DNA ploidy (PL) | S-phase (SP, %) |
|---|---|---|---|---|---|---|---|
| A. Male and female patients (learning set + first unknown test set) | | | | | | | |
| Male | 151 | 2.54 ± 0.23 | 3.47 ± 0.06 | 3.00 ± 0.14 | 1.33 ± 0.04 | 1.21 ± 0.03 | 8.40 ± 0.47 |
| Female | 140 | 3.07 ± 0.35 | 3.57 ± 0.08 | 3.35 ± 0.21 | 1.35 ± 0.04 | 1.18 ± 0.03 | 8.21 ± 0.42 |
| B. Survivors and nonsurvivors (learning set) | | | | | | | |
| Survivor | 120 | 1.52 ± 0.22 | 3.14 ± 0.07 | 2.32 ± 0.13 | 1.14 ± 0.03 | 1.06 ± 0.02 | 7.31 ± 0.36 |
| Nonsurvivor | 111 | 4.18 ± 0.32[b] | 3.92 ± 0.05[b] | 4.09 ± 0.18[b] | 1.56 ± 0.04[b] | 1.35 ± 0.04[b] | 9.39 ± 0.51[b] |

[a] Means ± SEM.
[b] $2p < 0.001$, $t$-test.

MELANOMA: CLASSIF1 TRIPLE MATRIX CLASSIFICATION

A.) RECLASSIFICATION OF LEARNING SET

| NR. | CLASSIFIER CATEGOR. | CATEGORY ABBREVIAT. | COIN | CLASSIFIER MASKS |
|---|---|---|---|---|
| 1 | SURVIVAL | A | 1.00 | 000 |
| 2 | DEATH | B | 1.00 | +++ |

| REC. NR. | DATAB: MELA6.BI4 RECORD LABELS | CLASSIF1-CLASSIFIC. | COIN FACT | SAMPLE CLASSIF.MASKS . = no value |
|---|---|---|---|---|
| 2 | 47888. A | A | 1.00 | 000 |
| 3 | 49896. A | A | .67 | 00+ |
| 4 | 51077. A | A | .67 | --+ |
| 6 | 58306. A | A | 1.00 | 000 |
| 7 | 63220. A | B | .67 | ++0 |
| 8 | 63231. A | A | 1.00 | 000 |
| 9 | 67200. A | A | 1.00 | --0 |
| 11 | 74733. A | B | .67 | ++0 |
| 12 | 76030. A | A | 1.00 | --- |
| 13 | 77598. A | B | .67 | ++- |
| 153 | 46931. B | B | 1.00 | +++ |
| 154 | 47071. B | A | 1.00 | 000 |
| 155 | 51680. B | A | 1.00 | 000 |
| 157 | 54535. B | B | .67 | ++0 |
| 158 | 70983. B | B | 1.00 | +++ |
| 159 | 77107. B | B | .67 | ++- |
| 160 | 77913. B | A | .67 | 0+0 |
| 162 | 76475. B | B | 1.00 | +++ |
| 163 | 53494. B | B | 1.00 | +++ |
| 164 | 71682. B | B | .67 | ++0 |

B) CLASSIFICATION OF UNKNOWN TEST SET

| | | | | |
|---|---|---|---|---|
| 1 | 59083. ? A | A | .67 | 00+ |
| 2 | 56627. ? A | A | 1.00 | 0-0 |
| 3 | 74007. ? A | A | 1.00 | 000 |
| 4 | 79931. ? A | A | 1.00 | 000 |
| 5 | 59731. ? A | A | 1.00 | 000 |
| 6 | 54563. ? A | A | .67 | -0+ |
| 7 | 53485. ? A | B | .67 | ++0 |
| 8 | 63089. ? A | A | 1.00 | --0 |
| 9 | 62437. ? A | B | 1.00 | +++ |
| 10 | 75891. ? A | A | 1.00 | 00- |
| 32 | 54748. ? B | B | 1.00 | +++ |
| 33 | 53400. ? B | B | .67 | ++- |
| 34 | 54520. ? B | B | .67 | ++- |
| 35 | 80046. ? B | B | 1.00 | +++ |
| 36 | 74562. ? B | B | 1.00 | +++ |
| 37 | 77415. ? B | B | 1.00 | +++ |
| 38 | 73113. ? B | A | .67 | 0+0 |
| 39 | 71988. ? B | B | 1.00 | +++ |
| 40 | 56694. ? B | B | 1.00 | +++ |
| 41 | 57738. ? B | A | .67 | 00+ |

pat-ID | truth (hidden truth during learning phase) | classification | classification coincidence factor (CCF)

**Table III**
**Asthma: Blood Clinical Chemistry**

| Clinical diagnosis | Number of patients (n) | CLASSIF1 classification (%)[a] Healthy | Asthmatic | Specificity/ sensitivity |
|---|---|---|---|---|
| A. Learning set | | | | |
| Healthy children | 18 | **100.0** | 0.0 | 100.0 |
| Asthmatic children | 40 | 30.0 | **70.0** | 70.0 |
| Negative/positive predictive values | | 60.0 | 100.0 | ARI 85.0 |
| B. Unknown test set | | | | |
| Healthy children | 6 | **66.7** | 33.3 | 66.7 |
| Asthmatic children | 10 | 30.0 | **70.0** | 80.0 |
| Negative/positive predictive values | | 57.1 | 77.7 | ARI 68.5 |

[a] 25–>30% optimized percentile thresholds (data bases GOERLI1.BI4/.BI6).

as compared to the learning set has a lower discriminatory potential with positive and negative predictive values of 77.7 and 57.1% (Table III).

The means of six [β-globulin, IgG3, complement CH100 titer, T4, and leukocyte (LKCS)] of the clinical chemistry parameters are significantly different between asthmatic and nonasthmatic children ($2p < 0.05$, $t$-test). Significant mean value differences do not necessarily parallel good discrimination potential. As shown in Table IV, only β-globulin has enough discriminatory potential for data pattern analysis. This is further substantiated by systematic analysis of the discriminatory potential of the statistically most significant single parameters such as β-globulin, by which asthmatic children are detectable with a sensitivity of 40.9% at 90% specificity for the recognition of healthy children, similarly as, for example, IgG3 (41.1%), β-globulin (40.9%), and LKCS (38.9%). Statistically different parameters may, however, also be low discriminating parameters, such as complement CH100 (22.5%) or T4 (19.0%). On the other hand, nonsignificantly different and low discriminating single parameters may prove quite useful in data pattern analysis such as ASAT (10.4%) (Table IV).

**Fig. 1** Melanoma classification for the known learning set (A) and the first unknown test set (B) of patients using the classifier of Table I. The classifications for the first 10 patients in each classification category are displayed. Patient group A represents 10-year survivors, while patients in group B did not survive. Patients are classified according to the best positional coincidence of the patient classification mask with one of the two classifier masks. The three selected classification parameters are tumor thickness (TD at position 1 of the classifier masks), the mean value of tumor thickness and tumor infiltration depth (TK at position 2), and percentage of S-phase tumor cells (SP at position 3). Classification is performed down to a CCF of ≥0.67, that is, to a positional coincidence of the patient classification mask with the best fitting classifier mask for two of the three classification parameters. The truth positions were left blank (?) for the test set patients (B) to make them invisible for the CLASSIF1 algorithm during the learning process.

**Table IV**
**Asthma: Single Parameter Sensitivity and CLASSIF1 Classificator Masks of Selected Clinical Chemistry Parameters**

| Classification parameters (selected from 49 parameters) | Single parameter sensitivity (%) at 90% specificity | Classification matrix | |
|---|---|---|---|
| | | N | A |
| 1. TRCS | 29.2 | 0 | − |
| 2. EOS | 30.9 | 0 | + |
| 3. ASAT | 10.4 | 0 | − |
| 4. TSH | 31.0 | 0 | − |
| 5. Ferritin | 23.3 | 0 | − |
| 6. IgE | 35.0 | 0 | + |
| 7. β-Globulin | 40.9 | 0 | − |

The classification (Table V) of manually evaluated flow cytometry histograms for lymphocyte relative cell frequency (%) and relative antigen expression of the 16 two-parameter immunophenotypes indicates ideal classification with 100.0% sensitivity/specificity and positive/negative predictive values for asthmatic and healthy children. This classification is reached not only for the known learning set (Table V), but equally for the unknown test set of patients (Table V); that is, the classification is robust. The discrimination was achieved with data from only four measurements (CD4/45RA, CD8/11b, CD21/19, and CD71/3) (Table VI), whereas the other 12 measurements provided less direct and redundant information that resulted in exclusion during the selection process.

The classification of the list mode files included only partially the same measurements as for the manually evaluated histograms. In particular the most

**Table V**
**Asthma: Flow Cytometry by Cell Frequency and Fluorescence Intensity**

| Clinical diagnosis | Pat. (*n*) | CLASSIF1 classification (%)[a] | | |
|---|---|---|---|---|
| | | Healthy | Asthmatic | Specificity/ sensitivity |
| A. Learning set | | | | |
| Healthy children | 19 | **100.0** | 0.0 | 100.0 |
| Asthmatic children | 39 | 0.0 | **100.0** | 100.0 |
| Negative/positive predictive values | | 100.0 | 100.0 | ARI 100.0 |
| B. Unknown test set | | | | |
| Healthy children | 6 | **100.0** | 0.0 | 100.0 |
| Asthmatic children | 9 | 0.0 | **100.0** | 100.0 |
| Negative/positive predictive values | | 100.0 | 100.0 | ARI 100.0 |

[a] 10–>15% optimized percentile thresholds (data bases GO5LEARN.BI4/.BI6).

**Table VI**
**Asthma: Selected Cell Frequency and Fluorescence Intensity Parameters[a]**

| Classification parameters (selected from 103 lymphocyte parameters) | Healthy ($n$ = 19) | Asthma ($n$ = 39) | Units | Classification matrix N | Classification matrix A |
|---|---|---|---|---|---|
| 1. CD45RA Ab on $CD4^-/CD45RA^+$ lymphocytes | 26.94 ± 2.94 | 103.38 ± 7.68[b] | Arbitrary units | 0 | + |
| 2. % $CD4^-/CD45RA^+$ lymphocytes | 11.57 ± 0.92 | 51.51 ± 1.90[b] | % of lymphocytes | 0 | + |
| 3. CD4Ab on $CD4^+/CD45RA^-$ lymphocytes | 115.31 ± 7.88 | 26.59 ± 1.60[b] | Arbitrary units | 0 | − |
| 4. % $CD4^+/CD45RA^-$ lymphocytes | 50.94 ± 2.51 | 14.17 ± 0.64[b] | % of lymphocytes | 0 | − |
| 5. % $CD5^+/CD19^+$ lymphocytes | 7.84 ± 0.86 | 8.49 ± 1.36[b] | Arbitrary units | 0 | + |
| 6. CD8 Ab on $CD8^+/CD11b^-$ lymphocytes | 67.16 ± 7.79 | 37.62 ± 3.61[b] | % of lymphocytes | 0 | − |
| 7. CD45RO Ab on $CD45RO^+/CD4^-$ lymphocytes | 41.36 ± 6.51 | 26.17 ± 4.35 | % of lymphocytes | 0 | − |

[a] Means ± SEM.
[b] $2p < 0.001$, $t$-test.

discriminant measurements such as CD4/45RA, CD8/11b, and CD21/19 were not available as list mode files. Nevertheless the same ideal 100.0% result for sensitivity/specificity, positive/negative predictive values was obtained by the exhaustive CLASSIF1 list mode evaluation of lympho-, mono-, and granulocyte parameters for the learning set (Table VII) as well as for the unknown test set patients (Table VII). A closer analysis of the selected classification parameters (Table VIII), shows that 7 (CD45/14, CD3/HLA-DR, CD4/29, CD71/3, CD3/

**Table VII**
**Asthma: Exhaustive Flow Cytometric List Mode Analysis**

| Clinical diagnosis | Pat. ($n$) | CLASSIF1 classification (%)[a] Healthy | Asthmatic | Specificity/ sensitivity |
|---|---|---|---|---|
| A. Learning set | | | | |
| Healthy children | 19 | **100.0** | 0.0 | 100.0 |
| Asthmatic children | 30 | 0.0 | **100.0** | 100.0 |
| Negative/positive predictive values | | 100.0 | 100.0 | ARI 100.0 |
| B. Unknown test set | | | | |
| Healthy children | 5 | **100.0** | 0.0 | 100.0 |
| Asthmatic children | 9 | 0.0 | **100.0** | 100.0 |
| Negative/positive predictive values | | 100.0 | 100.0 | ARI 100.0 |

[a] 10–>30% optimized percentile thresholds (data bases KHLEARN.BI4/.BI6).

**Table VIII**
**Asthma: Selected Parameters Exhaustive List Mode Analysis[a]**

| Classification parameters (selected from 759 lympho-, mono-, granulocyte parameters of 11 two color immunophenotypes) | References ($n$ = 19) | Asthma ($n$ = 30) | Units | Classification matrix N | Classification matrix A |
|---|---|---|---|---|---|
| 1. CD45Ab surf. dens.[b] on $CD45^+$ lymphocytes | 0.233 ± 0.007 | 0.170 ± 0.012[c] | Arbitrary units | 0 | − |
| 2. % $CD3^+$/HLA-$DR^+$ lymphocytes | 11.26 ± 1.58 | 3.66 ± 0.31[c] | % of lymphocytes | 0 | − |
| 3. HLA-DR Ab on $CD3^+$/HLA-$DR^+$ lymphocytes | 0.965 ± 0.091 | 1.971 ± 0.254[c] | Arbitrary units[d] | 0 | + |
| 4. CD4 Ab surf. dens. on $CD4^+$ lymphocytes | 0.132 ± 0.005 | 0.095 ± 0.006 | Arbitrary units | 0 | − |
| 5. % $CD71^+$ lymphocytes | 8.02 ± 1.24 | 3.63 ± 0.55[c] | % of lymphocytes | 0 | − |
| 6. CD3 Ab on $CD3^-$/$CD56^+$ granulocytes | 0.0894 ± 0.0046 | 0.0461 ± 0.0032 | Arbitrary units[d] | 0 | − |
| 7. CD56/CD3 Ab ratio on $CD3^-$/$CD56^+$ granulocytes | 4.30 ± 0.63 | 13.67 ± 1.88[c] | Arbitrary units | 0 | + |
| 8. CD4 Ab on $CD4^-$/$CD29^+$ granulocytes | 0.1000 ± 0.0054 | 0.0475 ± 0.0033[c] | Arbitrary units[d] | 0 | − |
| 9. % $CD25^+$ granulocytes | 78.28 ± 4.51 | 11.10 ± 2.93[c] | % of granulocytes | 0 | − |
| 10. CD19/CD5 Ab ratio on $CD5^-$/$CD19^+$ granulocytes | 1.66 ± 0.24 | 4.51 ± 1.76 | Arbitrary units | 0 | + |
| 11. % $CD5^+$/$CD19^-$ granulocytes | 42.80 ± 6.53 | 5.97 ± 2.38[c] | % of granulocytes | 0 | − |

[a] Means ± SEM.
[b] Ab surf. dens. = relative antibody surface density (linearized fluorescence/square root of forward light scatter).
[c] $2p < 0.001$, $t$-test.
[d] Arbitrary unit: 0.001–10 V fluorescence scale, relinearized from the four decade log fluorescence scale of the flow cytometer, ratio calculation from relinearized fluorescence values.

**Fig. 2** Juvenile asthma patient classification for the first 10 patients of the learning set (A) and the five and nine patients of the test set (B) using the classifier of Table VII. Eleven classification parameters were automatically selected by the CLASSIF1 algorithm from 759 data columns extracted from the lympho-, mono-, and granulocyte cell populations of 11 FITC/PE immunophenotype list mode files per patient (Table VIII). Parameters from seven immunophenotypes (CD45/14, CD3/HLA-DR, CD71/3, CD3/16+56, CD4/29, CD25/3, and CD5/19) are required for the classification. Classification is performed down to a CCF of ≥0.55, that is, to a positional coincidence of the classification mask of the patient with the best fitting classifier mask for 6 of the 11 classification parameters. Patient 0001 of the test set is not classified (−) because of a CCF of 0.27 as a consequence of missing values (.) by nonavailable list mode files. Due to coincidence for three positions with the asthma classifier mask (A) and two with the mask of normal children (N), the patient was manually classified as asthma for the test set classification (Table VII). The truth position for all test set patients were left blank (?) during the learning phase as in Fig. 1.

JUVENILE ASTHMA: CLASSIF1 TRIPLE MATRIX CLASSIFICATION

A.) RECLASSIFICATION OF LEARNING SET

| NR. | CLASSIFIER CATEGOR. | CATEGORY ABBREVIAT. | COIN | CLASSIFIER MASKS |
|---|---|---|---|---|
| 1 | NORMAL | N | 1.00 | 00000000000 |
| 2 | ASTHMA | A | 1.00 | --+---+--+- |

| REC. NR. | DATAB: KHLEARN.BI4 RECORD LABELS | CLASSIF1-CLASSIFIC. | CLAS COIN | SAMPLE CLASSIF.MASKS . = no value |
|---|---|---|---|---|
| 40 | 0059 N | N | 1.00 | +00++0-++-+ |
| 41 | 0058 N | N | .64 | 00+000+0-0- |
| 42 | 0061 N | N | 1.00 | 0+00000+000 |
| 43 | 0060 N | N | .82 | 0-000000-00 |
| 44 | 0064 N | N | .73 | 000+000--0- |
| 45 | 0066 N | N | 1.00 | 0000+000000 |
| 46 | 0065 N | N | .55 | 0-00-00-0+- |
| 48 | 0067 N | N | 1.00 | 00-00+0+0-+ |
| 49 | 0069 N | N | .64 | 0++0000-++- |
| 50 | 0070 N | N | .82 | 0-00-00000+ |
| 2 | 0003 A | A | .55 | +-0+-00--+- |
| 3 | 0004 A | A | .64 | +-++00+--+- |
| 5 | 0008 A | A | .82 | 0-++--+--+- |
| 6 | 0009 A | A | 1.00 | --+---+--+- |
| 7 | 0010 A | A | 1.00 | --+---+--+- |
| 9 | 0014 A | A | .91 | --+-+-+--+- |
| 10 | 0015 A | A | 1.00 | --+---+--+- |
| 12 | 0018 A | A | 1.00 | --+---+--+- |
| 13 | 0027 A | A | .91 | --+--0+--+- |
| 14 | 0030 A | A | 1.00 | --+---+--+- |

B) CLASSIFICATION OF UNKNOWN TEST SET

| | | | | | |
|---|---|---|---|---|---|
| 47 | 0068 ? | N | N | .91 | 0+-+0+-++-- |
| 53 | 0073 ? | N | N | .64 | 0-+0000-0++ |
| 58 | 0078 ? | N | N | .82 | 00-+0000-0- |
| 60 | 0080 ? | N | N | .55 | 0-+-000-0++ |
| 61 | 0083 ? | N | N | 1.00 | 0+00000+0-+ |
| 1 | 0001 ? | A | - | .27 | --+..00.... |
| 4 | 0007 ? | A | A | .82 | +-++--+--+- |
| 8 | 0012 ? | A | A | .82 | --+---0-0+- |
| 11 | 0017 ? | A | A | 1.00 | --+---+--+- |
| 16 | 0032 ? | A | A | .64 | --+-.-+-... |
| 21 | 0037 ? | A | A | 1.00 | --+---+--+- |
| 26 | 0042 ? | A | A | .91 | -0+---+--+- |
| 31 | 0048 ? | A | A | .91 | --+0--+--+- |
| 35 | 0053 ? | A | A | .82 | --+-0-+-0+- |

pat-ID | truth | hidden truth during learning phase | classification | classification coincidence factor (CCF)

**Table IX**
**Asthma: Exhaustive List Mode Analysis of CD25/3**

| Clinical diagnosis | Number of patients (*n*) | CLASSIF1 classification (%)[a] Healthy | Asthmatic | Specificity/ sensitivity |
|---|---|---|---|---|
| A. Learning set | | | | |
| Healthy children | 20 | **100.0** | 0.0 | 100.0 |
| Asthmatic children | 31 | 3.2 | **96.8** | 96.8 |
| Negative/positive predictive values | | 95.2 | 100.0 | ARI 98.4 |
| B. Unknown test set | | | | |
| Healthy children | 4 | **100.0** | 0.0 | 100.0 |
| Asthmatic children | 8 | 12.5 | **87.5** | 87.5 |
| Negative/positive predictive values | | 80.0 | 100.0 | ARI 93.7 |

[a] 10–>15% optimized percentile thresholds (data bases KQLEARN.BI4/.BI6).

56, CD25/3, and CD5/19) of the 11 measurements were required for classification. The printout of the classification masks (Fig. 2) for the individual patients of the learning and test sets indicates robustness of classification in case of partial nonidentity between the patient classification mask and the best fitting classifier mask. Minimally 6 of the 11 classifier parameters have to match with the selected classifier mask to avoid sample rejection at the observed CCF of 0.55.

The separate classification of each individual two-parameter immunophenotype with simultaneous consideration of lympho-, mono-, and granulocyte data shows that, for example, the analysis of the single CD25/3 immunophenotype alone discriminates already quite well (ARI = 98.4%) between asthmatic and nonasthmatic children in the learning set (Table IX) as well as in the unknown test set patients (Table IX) with a selection of three cell frequency parameters (Table X). Similar results were obtained for CD57/8 (ARI = 95.9%) and CD5/19 (ARI = 95.9%).

**Table X**
**Asthma: Selected Parameters Exhaustive CD25/3 Analysis[a]**

| Classification parameters (selected from 69 lympho-, mono-, and granulocyte parameters) | Healthy ($n$ = 20) | Asthma ($n$ = 31) | Classification matrix N | A |
|---|---|---|---|---|
| 1. % $CD25^+$ granulocytes | 78.28 ± 4.51 | 11.10 ± 2.93[b] | 0 | – |
| 2. % $CD25^-/CD3^+$ granulocytes | 0.224 ± 0.068 | 3.41 ± 1.19[b] | 0 | + |
| 3. % $CD25^+/CD3^+$ granulocytes | 9.47 ± 3.16 | 2.65 ± 1.19[b] | 0 | – |

[a] Means ± SEM (% of granulocytes).
[b] $2p < 0.05$, $t$-test.

**Table XI**
**HIV Infection: Flow Cytometry by Cell Frequency**

| Clinical diagnosis | Number of patients ($n$) | CLASSIF1 classification (%)[a] Seronegative | Seropositive | Specificity/ sensitivity |
|---|---|---|---|---|
| A. Learning set | | | | |
| Seronegative | 15 | **100.0** | 0.0 | 100.0 |
| Seropositive | 55 | 7.3 | 92.7 | 82.6 |
| Negative/positive predictive values | | 78.9 | **100.0** | ARI 96.4 |
| B. Unknown test set | | | | |
| Seronegative | 5 | **100.0** | 0.0 | 100.0 |
| Seropositive | 14 | 0.0 | **100.0** | 100.0 |
| Negative/positive predictive values | | 100.0 | 100.0 | ARI 100.0 |

[a] 10–90% percentile thresholds (data base CD26TOT6.BI4/.BI6).

## C. HIV Infection

The classification (Table XI) of the 18 parameters from manual analysis of two-color lymphocyte immunophenotype histograms, including the white blood cell and lymphocyte counts (WBC, LYC), provides positive and negative predictive values of 100.0 and 78.9% for HIV seropositive and seronegative patients with similar values for the unknown test set patients (Table XI). The selected three parameters (Table XII) involve CD45RA/4, HLA-DR/CD8, and CD8/38 immunophenotype measurements.

The exhaustive lympho-, mono-, and granulocyte parameter extraction by the CLASSIF1 analysis provided average recognition between 96.1 and 100.0% (ARI) at multiplicity indices between 1.00 and 1.02 (AMI) for the individual evaluation of either the CD2/19, HLA-DR/CD8, CD45RA/4, or the CD8/38 measurement. Evaluation of only the lymphocyte cell population provided ARIs

**Table XII**
**HIV Infection: Selected Lymphocyte Frequency Parameters[a]**

| Classification parameters (selected from 18 lymphocyte/ leukocyte parameters) | Seronegative ($n$ = 15) | Seropositive ($n$ = 55) | Units | Classification matrix N | P |
|---|---|---|---|---|---|
| 1. $CD45RA^+/CD4^+$ lymphocytes | 23.53 ± 1.94 | 8.14 ± 0.83[b] | % of lymphocytes | 0 | − |
| 2. $HLA\text{-}DR^+/CD8^+$ lymphocytes | 7.06 ± 1.04 | 36.01 ± 1.86[b] | % of lymphocytes | 0 | + |
| 3. $CD8^+/CD38^+$ lymphocytes | 13.60 ± 1.04 | 45.01 ± 2.30[b] | % of lymphocytes | 0 | + |

[a] Means ± SEM.
[b] $2p < 0.001$, $t$-test.

**Table XIII**
**HIV Infection: Exhaustive Flow Cytometric List Mode Analysis on Lymphocytes (HLA-DR/CD8)**

| Clinical diagnosis | Number of patients ($n$) | CLASSIF1 classification (%)[a] Seronegative | Seropositive | Specificity/ sensitivity |
|---|---|---|---|---|
| A. Learning set | | | | |
| Seronegative | 15 | **100.0** | 0.0 | 100.0 |
| Seropositive | 55 | 0.0 | **100.0** | 100.0 |
| Negative/positive predictive values | | 100.0 | 100.0 | ARI 100.0 |
| B. Unknown test set | | | | |
| Seronegative | 5 | **100.0** | 0.0 | 100.0 |
| Seropositive | 14 | 0.0 | **100.0** | 100.0 |
| Negative/positive predictive values | | 100.0 | 100.0 | ARI 100.0 |

[a] 15–85% percentile thresholds (data base PRLEARN.BI4/.BI6).

of 100.0% for HLA-DR/CD8 (Table XIII), 99.3% for CD45RA/4, and 97.8% for CD8/38, all at 1.00 multiplicity. HLA-DR/CD8 provided in addition positive and negative predictive values of 100% for the HIV seronegative and the seropositive patients in various disease states (seronegative $n = 15$, seropositive WHO stage 1/2/3/4 $n = 21/9/14/11$) for the learning set (Table XIII) as well as in the test set patients (Table XIII) ($n = 5/5/2/3/4$). Four of the five selected parameters concern antigen expression, relative antigen surface density, and antigen ratios and only one concerns a percent cell frequency parameter (Table XIV). The listing of the triple matrices for the individual patients (Fig. 3) indicates robustness of classification for some degree of positional nonidentity between the patient classi-

**Table XIV**
**HIV Infection: Selected HLA-DR/CD8 Lymphocyte Parameters[a]**

| Classification parameters (selected from 22 lymphocyte parameters | Seronegative ($n = 15$) | Seropositive ($n = 55$) | Units | Classification matrix N | P |
|---|---|---|---|---|---|
| 1. CD8 Ab on $CD8^+$ lymphocytes | 18.25 ± 1.02 | 7.91 ± 0.23[b] | Arbitrary units | 0 | – |
| 2. CD8 rel. Ab surf. dens. on $CD8^+$ lymphocytes | 0.761 ± .046 | 0.328 ± 0.009[b] | Arbitrary units | 0 | – |
| 3. CD8 Ab on $HLA\text{-}DR^-/CD8^+$ lymphocytes | 18.01 ± 1.04 | 7.63 ± 0.22[b] | Arbitrary units | 0 | – |
| 4. CD8/HLA-DR Ab ratio on $HLA\text{-}DR^-/CD8^+$ lymphocytes | 671.1 ± 47.7 | 197.3 ± 10.5[b] | Arbitrary units | 0 | – |
| 5. % $HLA\text{-}DR^+/CD8^+$ lymphocytes | 3.68 ± 0.51 | 24.12 ± 1.44[b] | % of lymphocytes | 0 | + |

[a] Means ± SEM.
[b] $2p < 0.001$, $t$-test.

HIV INFECTION: CLASSIF1 TRIPLE MATRIX CLASSIFICATION

A.) RECLASSIFICATION OF LEARNING SET

| NR. | CLASSIFIER CATEGOR. | CATEGORY ABBREVIAT. | COIN | CLASSIFIER MASKS |
|---|---|---|---|---|
| 1 | NORMAL | N | 1.00 | 00000 |
| 2 | SEROPOS | P | 1.00 | ----+ |

| REC. NR. | DATAB: PRLEARN.BI4 RECORD LABELS | | CLASSIF1-CLASSIFIC. | CLAS COIN | SAMPLE CLASSIF.MASKS . = no value |
|---|---|---|---|---|---|
| 55 | KE0K02 | N | N | 1.00 | 00000 |
| 56 | KE0K03 | N | N | 1.00 | 0000- |
| 57 | KE0K04 | N | N | .60 | 000-+ |
| 58 | KE0K06 | N | N | .80 | +++++ |
| 59 | KE0K07 | N | N | .80 | +++++ |
| 60 | KE0K08 | N | N | 1.00 | ++++0 |
| 61 | KE0K09 | N | N | 1.00 | 00000 |
| 62 | KE0K11 | N | N | 1.00 | 00000 |
| 63 | KE0K12 | N | N | .60 | -0-00 |
| 1 | KE3756 | P | P | 1.00 | ----+ |
| 2 | KE3758 | P | P | 1.00 | ----+ |
| 3 | KE3759 | P | P | 1.00 | ----+ |
| 4 | KE3766 | P | P | 1.00 | ----+ |
| 5 | KE3767 | P | P | 1.00 | ----+ |
| 6 | KE3768 | P | P | 1.00 | ----+ |
| 7 | KE3769 | P | P | 1.00 | ----+ |
| 8 | KE3771 | P | P | 1.00 | ----+ |
| 9 | KE3772 | P | P | 1.00 | ----+ |
| 10 | KE3773 | P | P | 1.00 | ----+ |

B) CLASSIFICATION OF UNKNOWN TEST SET

| | | | | | |
|---|---|---|---|---|---|
| 87 | KE0K01 | ? N | N | 1.00 | 00000 |
| 91 | KE0K05 | ? N | N | .80 | 0000+ |
| 96 | KE0K10 | ? N | N | 1.00 | 000+- |
| 101 | KE0K15 | ? N | N | 1.00 | 00000 |
| 106 | KE0K20 | ? N | N | 1.00 | 00000 |
| 17 | KE3755 | ? P | P | 1.00 | ----+ |
| 21 | KE3765 | ? P | P | 1.00 | ----+ |
| 26 | KE3770 | ? P | P | 1.00 | ----+ |
| 31 | KE3776 | ? P | P | .80 | ----0 |
| 36 | KE3782 | ? P | P | 1.00 | ----+ |
| 41 | KE3789 | ? P | P | 1.00 | ----+ |
| 46 | KE3795 | ? P | P | 1.00 | ----+ |
| 52 | KE3993 | ? P | P | 1.00 | ----+ |
| 57 | KE4001 | ? P | P | 1.00 | ----+ |
| 63 | KE4061 | ? P | P | 1.00 | ----+ |

pat-ID | truth hidden truth during learning phase | classification | classification coincidence factor (CCF)

**Fig. 3** Classification of the learning (A) and test set (B) of HIV seronegative (N) and seropositive (P) patients using the classifier of Table XIII. Five parameters per patient were selected from 22 lymphocyte data columns of HLA-DR/CD8 immunophenotype list mode analysis (Table XIV). Classification is performed down to a CCF of $\geq 0.60$, that is, to a positional coincidence for three of the five classification parameters of the patient's classification mask with the best fitting classifier mask.

fication mask and the best fitting classifier mask. A minimum of three positional coincidences with the selected five parameters classification matrix is required (CCF = 0.60).

## IV. Discussion

The three classification examples from unrelated clinical areas show the potential of data pattern classification for disease course prediction (melanoma) as well as for a precise biomolecular diagnosis (juvenile asthma, HIV infection). Precise diagnosis represents a precondition for the elaboration of predictive classifiers.

The predictive capacity of the melanoma classifier (Table I) is similar to the one for survival prediction in colorectal carcinoma patients (van Driel *et al.*, 1999) that is, lower than for the estimation of sepsis outcome in intensive care medicine (Rothe *et al.*, 1990; Valet *et al.*, 1998) as well as for the preoperative prediction of postcardiotomy syndrome in children with open heart surgery (Tarnok *et al.*, 1997, 1999). This is caused by the comparatively small initial parameter pattern of four clinical and only two flow cytometric parameters (Table II). In spite of the few parameters, the classification of the 231 learning set patients and especially of the 268 unknown test patients provides stable results.

When comparing the information content in the various measurements in juvenile asthma and HIV infected patients, only a small fraction of the available parameters contains the discriminant information. In juvenile asthma, the diagnostic information is encountered in 14.2% (7 of 49) of the clinical chemistry parameters (Table IV), in 6.7% (7 of 103) of the lymphocyte analysis (Table VI) provided by 4 of 12 FITC/PE immunophenotypes, in 1.4% (11 of 759) of the parameters from exhaustive lympho-, mono-, and granulocyte analysis from 7 of 11 FITC/PE immunophenotypes (Table VIII), and in 4.3% (3 of 69) of the parameters from single CD25/3 immunophenotype analysis (Table X). A similar situation was encountered in the analysis of HIV infected patients (16.6%, 22.7%, Tables XII and XIV).

The confinement of discrimination to relatively few biomolecular parameters was similarly encountered in intensive care medicine for the determination of cell function parameters (Valet *et al.*, 1993; Rothe *et al.*, 1990) as well as for immunophenotyping, in particular in lymphomas (Valet and Höffkes, 1997), in the expression of thrombocyte surface antigens for myocardial infarction risk assessment (Valet *et al.*, 1993), in the prediction of the postcardiotomy syndrome in children (Tarnok *et al.*, 1997, 1999), and in the early detection of the overtraining syndrome in competition cyclists (Gabriel *et al.*, 1993, 1998; Valet *et al.*, 1993).

Considering the diversity of these diseases, the restriction of discrimination to a relatively narrow biomolecular parameter pattern seems to represent a more general rule. It comprises the potential for a significantly higher impact of predictive and diagnostic achievements for the individual patient at equal efforts. The advantage of data pattern analysis is that the discriminatory data pattern is

provided in a standardized way, accessible to international efforts of consensus formation and optimization as evidence based medicine (EBM) at a cellular level.

The results of the immunophenotype classifications in asthma (Tables VIII and X) reemphasize the earlier observation of lymphoma immunophenotyping (Valet and Höffkes, 1997) showing that a significant amount of discriminatory information is localized on granulocytes or monocytes although the antibody panels are primarily selected for lymphocyte antigens. The reason for this seems to be either reactive adaptation of existing nonlymphocytic cell populations to the disease process or a reactively altered formation of cell populations by the hemopoietic organs.

Concerning the issue of whether the evaluation of percent cell frequency parameters is sufficient or whether the more complex quantitative analysis of antibody binding is required, Tables VI, VIII, X, and XIV clearly show that a substantial number of the discriminatory parameters are antibody intensity and antibody binding ratios. It seems therefore mandatory to routinely evaluate fluorescence intensities, fluorescence ratios, and in the future also coefficients of variations for all cell population parameters in flow cytometric histograms.

Backed by the information provided in this chapter and from earlier results, it seems clear that exhaustive information extraction from clinical multiparameter flow cytometry measurements in combination with discriminant data pattern analysis will constitute an important access route for disease course prediction at the individual patient level. Although the currently presented classification work concerns retrospectively prospective metaanalysis, it can be reasonably assumed that the classifiers will perform equally well in prospective studies. This hope is deduced from the observed robustness of classification of unknown samples in all the various studies performed up to now with the CLASSIF1 algorithm.

## References

Beckman, R. J., Salzman, G. C., and Stewart, C. C. (1995). Classification and regression trees for bone marrow immunophenotyping. *Cytometry* **20,** 210–217.

Boddy, L., Morris, C. W., Wilkens, M. F., Tarran, G. A., and Burkill, P. H. (1994). Neural network analysis of flow cytometric data for 40 marine phytoplankton species. *Cytometry* **15,** 283–293.

Davey, H. M., Jones, A., Shaw, A. D., and Kell, D. B. (1999). Variable selection and multivariate methods for the identification of micrororganisms by flow cytometry. *Cytometry* **35,** 162–168.

Decaestecker, C., Remmelink, M., Salmon, I., Camby, I., Goldschmidt, D., Patein, M., Van Ham, P., Pasteels, J. L., and Kiss, R. (1996). Methodological aspects of using decision trees to characterise Leiomyomatous tumors. *Cytometry* **24,** 83–92.

Demers, S., Kim, J., Legendre, P., and Legendre, L. (1992). Analyzing multivariate flow cytometric dara in aquatic sciences. *Cytometry* **13,** 291–298.

Diamond, L. W., Nguyen, D. T., Andreeff, M., Maiese, R. L., and Braylan, R. C. (1994). A knowledge-based system for the interpretation of flow cytometric data in leukemia and lymphomas. *Cytometry* **17,** 266–273.

Frankel, D. S., Olsen, R. J., Frankel, S. I., and Chisholm, S. W. (1989). Use of a neural net computer system for analysis of flow cytometric data of phytoplankton populations. *Cytometry* **10,** 540–550.

Frankel, D. S., Frankel, S. L., Binder, B. J., and Vogt, R. F. (1996). Application of neural networks to flow cytometry data analysis and real-time cell classification. *Cytometry* **23,** 290–302.

Gabriel, H., Valet, G., Urhausen, A., and Kindermann, W. (1993). Selbstlernende Klassifizierung durchfluβzytometrischer Listendaten von immunphänotypisierten Lymphozyten bei akuter körperlicher Arbeit. *Deutsche Zschr. Sportmedizin* **44,** 461–465.

Gabriel, H., Urhausen, A., Valet, G., Heidelbach, U., and Kindermann, W. (1998). Overtraining and immune system: A prospective longitudinal study in endurance athletes. *Med. Sci. Sports Exerc.* **30,** 1151–1157.

Hokanson, J. A., Rosenblatt, J. I., and Leary, J. F. (1999). Some theoretical and practical considerations for multivariate statistical cell classification useful in autologous stem cell transplantation and tumor cell purging. *Cytometry* **36,** 60–70.

Molnar, B., Szentirmay, Z., Bodo, M., Sugar, J., and Feher, J. (1993). Application of multivariate, fuzzy set and neural network analysis in quantitative cytological examinations. *Anal. Cell. Pathol.* **5,** 161–175.

Leary, J. F. (1994). Strategies for rare cell detection and isolation. *In* "Methods in Cell Biology: Flow Cytometry" (Z. Darzynkiewicz, J. P. Robinson, and H. A. Crissman, eds.), 2nd Ed., Part B, Vol. 42. pp. 331–358. Academic Press, San Diego.

Otto, F. J., Oldiges, H., Göhde, W., and Jain, V. K. (1981). Flow cytometric measurementy of nuclear DNA content variation as a potential in vivo mutagenicity test. *Cytometry* **2,** 188–191.

Ravdin, P. M., Clark, G. M., Hough, J. J., Owens, M. A., and McGuire, W. L. (1993). Neural network analysis of DNA flow cytometry histograms. *Cytometry* **14,** 74–80.

Rothe, G., Kellermann, W., and Valet, G. (1990). Flow cytometric parameters of neutrophil function as early indicators of sepsis or trauma-related pulmonary or cardiovascular failure. *J. Lab. Clin. Med.* **115,** 52–61.

Schut, T. C. B., De Grooth, B. G., and Greve, J. (1993). Cluster analysis of flow cytometric list mode data on a personal computer. *Cytometry* **14,** 649–659.

Tarnok, A., Hambsch, J., Borte, M., Valet, G., and Schneider, P. (1997). Immunological and serological discrimination of children with and without post-surgical capillary leak syndrome. *In* "The Immune Consequences of Trauma, Shock and Sepsis" (E. Faist, ed.), pp. 845–849. Monduzzi Editore, Bologna.

Tarnok, A., Pipek, M., Valet, G., Richter, J., Hambsch, J., and Schneider, P. (1999). Children with post-surgical capillary leak syndrome can be distinguished by antigen expression on neutrophils and monocytes. *In* "Progress in Biomedical Optics, Proceedings Systems and Technologies for Clinical Diagnostics and Drug Discovery II" (G. E. Cohn and J. C. Owicki, eds.), SPIE Vol. 3603, pp. 61–71. Int. Soc. for Optical Engineering, Bellingham, WA.

Terstappen, L. W. M., Mickaels, R. A., Dost, R., and Loken, M. R. (1990). Increased light scattering resolution facilitates multidimensional flow cytometric analysis. *Cytometry* **11,** 506–512.

Thews, O., Thews, A., Huber, C., and Vaupel, P. (1996). Computer-assisted interpretation of flow cytometry data in hematology. *Cytometry* **23,** 140–149.

Valet, G., and Höffkes, H. G. (1997). Automated classification of patients with chronic lymphatic leukemia and immunocytoma from flow cytometric three colour immunophenotypes. *Cytometry (Commun. Clin. Cytometry)* **30,** 275–288.

Valet, G., Valet, M., Tschöpe, D., Gabriel, H., Rothe, G., Kellermann, W., and Kahle, H. (1993). White cell and thrombocyte disorders: Standardized, self-learning flow cytometric list mode data classification with the CLASSIF1 program system. *Ann. N.Y. Acad. Sci.* **677,** 233–251.

Valet, G., Roth, G., and Kellermann, W. (1998). Risk assessment for intensive care patients by automated classification of flow cytometric oxidative burst, serine and cysteine proteinase activity measurements using CLASSIF1 triple matrix analysis. *In* "Cytometric Cellular Analysis" (J. P. Robinson and G. Babcock, eds.), pp. 289–306. Wiley-Liss, New York.

Van Driel, B. E. M., Valet, G. K., Lyon, H., Hansen, U., Song, J. Y., Van Noorden, C. J. F. (1999). Prognostic estimation of survival of colorectal cancer patients with the quantitative histochemical assay of G6PDH activity and the multiparameter classification program CLASSIF1. *Cytometry (Commun. Clin. Cytometry)* **38,** 176–183.

Verwer, B. J. H., and Terstappen, L. W. M. M. (1993). Automatic lineage assignment of acute leukemias by flow cytometry. *Cytometry* **14,** 862–875.