Cytometry
PART A
Journal of the International Society for Analytical Cytology

Automated Classification of Bacterial Particles in Flow by Multiangle Scatter Measurement and Support Vector Machine Classifier

Bartek Rajwa,[1]* Murugesan Venkatapathi,[1,2] Kathy Ragheb,[1] Padmapriya P. Banada,[3] E. Daniel Hirleman,[2] Todd Lary,[4] J. Paul Robinson[1]

[1]Purdue University Cytometry Laboratories, Bindley Bioscience Center, Purdue University, West Lafayette, Indiana 47907

[2]School of Mechanical Engineering, Purdue University, West Lafayette, Indiana 47907

[3]Molecular Food Microbiology Laboratory, Department of Food Science, Purdue University, West Lafayette, Indiana 47907

[4]Cellular Analysis Technology Center, Beckman-Coulter, Inc., Miami, Florida 33196

Received 29 August 2007; Accepted 16 November 2007

*Correspondence to: Bartek Rajwa, Purdue University Cytometry Laboratories, Bindley Bioscience Center, Purdue University, 1203 W. State Street, West Lafayette, IN 47907, USA.

Email: brajwa@purdue.edu

Published online 28 December 2007 in Wiley InterScience (www.interscience.wiley.com)

DOI: 10.1002/cyto.a.20515

• **Abstract**

Biological microparticles, including bacteria, scatter light in all directions when illuminated. The complex scatter pattern is dependent on particle size, shape, refraction index, density, and morphology. Commercial flow cytometers allow measurement of scattered light intensity at forward and perpendicular (side) angles ($2^\circ \leq \theta_1 \leq 20^\circ$ and $70^\circ \leq \theta_2 \leq 110^\circ$, respectively) with a speed varying from 10 to 10,000 particles per second. The choice of angle is dictated by the fact that scattered light in the forward region is primarily dependent on cell size and refractive index, whereas side-scatter intensity is dependent on the granularity of cellular structures. However, these two-parameter measurements cannot be used to separate populations of cells of similar shape, size, or structure. Hence, there have been several attempts in flow cytometry to measure the entire scatter patterns. The published concepts require the use of unique custom-built flow cytometers and cannot be applied to existing instruments. It was also not clear how much information about patterns is really necessary to separate various populations of cells present in a given sample. The presented work demonstrates application of pattern-recognition techniques to classify particles on the basis of their discrete scatter patterns collected at just five different angles, and accompanied by the measurement of axial light loss. The proposed approach can be potentially used with existing instruments because it requires only the addition of a compact enhanced scatter detector. An analytical model of scatter of laser beams by individual bacterial cells suspended in a fluid was used to determine the location of scatter sensors. Experimental results were used to train the support vector machine-based pattern recognition system. It has been shown that information provided just by five angles of scatter and axial light loss can be sufficient to recognize various bacteria with 68–99% success rate. © 2007 International Society for Analytical Cytology

• **Key terms**

multiangle light-scatter; support vector machines; machine learning; pattern recognition; discrete-dipole approximation

Light-scatter signal detection has been employed in flow cytometry almost from the moment the method was introduced to practical use. Initially, scatter signal was utilized to synchronize the fluorescence detectors with the flow of particles through the flow chamber. Very soon it was demonstrated that information about forward ($2^\circ \leq \theta_1 \leq 20^\circ$) and side ($70^\circ \leq \theta_2 \leq 110^\circ$) scatter can be used to identify a number of subpopulations of cells without the use of any additional information provided by fluorescence stains (1–3). This was possible owing to the fact that forward-scattered light in the small-angle region ($\theta \leq 2^\circ$) is primarily dependent on the cell size, and is mostly independent of particle refractive index or shape (4–6), whereas perpendicular light scatter is sensitive to small internal structures in cells and to refractive index changes. In the early days of flow cytometry there had been also some reports published on the use of axial light loss, which was employed for cell sizing (7).

ISAC
International Society for Analytical Cytology

Flow cytometrists agree that full scatter patterns of bioparticles may contain much more information than what forward scatter, perpendicular (side) scatter, and extinction can reveal. When a cell passing through the flow chamber is illuminated, a very complex spatial pattern is formed that is dependent on cell size, shape, refraction index, density, morphology, and orientation of the cell relative to direction of incident beam. Therefore, many researchers investigated the possibility of collecting more than just two angles of scatter simultaneously. Meyer et al. (8) postulated that single cells could be comprehensively characterized in flow systems using multiangle scatter detectors utilizing 32 channels in a fashion similar to the observation of scatter patterns of single cells in microscopy-based systems.

Despite technical difficulties there were indeed several reports published in the 1970s and early 1980s on complex applications of scatter detection in flow, such as label-free detection of morphological changes inside the cell. Some of the reported systems involved detection of the full 180° or 360° scatter patterns from single biological cells (1,9–11).

The largest body of work on scatter in flow cytometry was performed in the 1970s by a group of researchers at Los Alamos National Labs (12–14). Their custom-built flow cytometers capable of multiangle scatter measurement were interfaced to DEC minicomputers for data processing. Several of these instruments were delivered to the NIH, but sadly none of the multiangle scatter detectors designed in Los Alamos found its way to commercial systems.

Owing to the immense progress in fluorescent label development, multiparameter flow cytometry moved in the last decade in the direction of adding more fluorescence detectors, rather than enhancing scatter measurements. Twelve-color machines are currently commercially available, and reports have been published on 16-color flow cytometry analysis. Multiangle scatter systems designed in the 1970s and 1980s failed to make a substantial impact on the field. Currently, only a few research groups still actively investigate applications of multiangle light-scatter analysis in flow (15–19). However, the published concepts usually require sophisticated, unique custom-built flow cytometers and cannot be applied to the existing instruments.

The presented work demonstrates application of pattern recognition techniques to classification of microbial particles on the basis of their discrete scatter patterns collected at five selected angles and accompanied by the measurement of axial light loss. Our approach differs from previous reports by the use of a discrete-dipole approximation (DDA)-based analytical model of laser-beam scatter by individual bacterial cells suspended in a fluid to determine the optimal location of scatter sensors. Most of the available cytometry literature related to light-scatter measurements employed generalized Lorenz–Mie theory to perform optical particle sizing using a model of light scattering by a sphere irradiated by a laser beam having a Gaussian intensity distribution. In contrast, the DDA method can be applied to nonhomogenous particles of arbitrary geometry, which makes it especially well suited for modeling scatter response of nonspherical cells, such as rod-shaped bacteria (20–22).

Our approach can be used with existing instruments and requires only minor modifications and the addition of a compact custom-built scatter detector. In contrast to other reports that describe direct use of collected scatter signals to characterize bioparticles, our method works in concert with a machine learning system. The experimental results obtained from the known samples are used to train the support vector machine (SVM), and subsequently the samples containing mixture of unknown particles are classified by the trained algorithm. This report shows that information provided by just six scatter-related parameters is sufficient for a trained system to recognize various bacteria with a 69–99% success rate.

Materials and Methods

Flow Cytometry

All the analyses were performed with a Cytomics FC500 flow cytometer (Beckman-Coulter, Miami, FL) equipped with a 488-nm air-cooled argon laser. A prototype of an enhanced scatter detection system (courtesy of Beckman-Coulter) capable of measuring forward-scatter signals at four different angles was added to the above flow cytometer, replacing the traditional forward scatter detector. This scatter measurement system consists of four ring detectors and an axial light-loss (extinction cross section) detector that can be moved toward or away from the laser beam-particle intersection point to change the angles of measurement (Fig. 1). Therefore, the four angles of detection in each experiment cannot be chosen independently because the four rings in the detector are fixed with respect to each other. The number of the uniformly spaced optical fibers in each ring varies linearly with its radius (12–34 per ring) to correct for variation of solid angle. The scatter measurements from each ring detector are amplified by different sets of pre-amplifiers and amplifiers to collect the 10-bit linear data. Discrimination of doublets from single particles is achieved by plotting forward-angle light-scatter integral versus peak intensity and gating on single-particle signals. The CXP software (Beckman-Coulter, Miami, FL) was used to acquire the data on the flow cytometer.

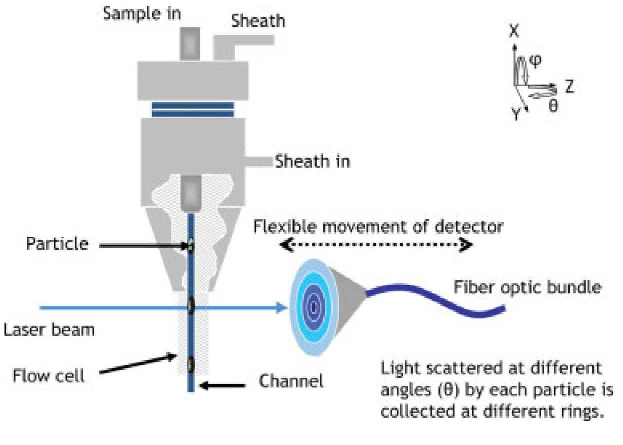
Sample in
Sheath
X
φ
Z
Y
θ
Sheath in
Flexible movement of detector
Particle
Fiber optic bundle
Laser beam
Flow cell
Channel
Light scattered at different angles (θ) by each particle is collected at different rings.

Figure 1. A simple schematics of the optical setup. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

Bacterial Cultures

Four different nonpathogenic bacterial cultures of varying size and shape were selected for the experiments: *Escherichia coli* K12, *Listeria innocua* F4248, *Bacillus subtilis* ATCC 6633, and *Enterococcus faecalis* CG110. The cultures were grown in brain heart infusion (BHI) broth for 16–18 h at 37°C, 140 rpm in a shaker incubator. The cultures were washed once by centrifuging 5 min at 3,000 rpm and resuspended in sterile phosphate buffered saline (PBS), pH 7.6, before analysis. All the bacterial cultures were obtained from the Purdue Department of Food Science culture collection.

Analytical Model of Scatter

The mathematical model of scatter used in this work assumes that the particles are in isolation in the sheath fluid ($n = 1.33$), and the angular scatter distribution is calculated and integrated over the area of the forward-scatter detector placed outside the sheath fluid. This assumption is valid if the particles are much smaller than the channel and if the laser beam (10 μm × 80 μm Gaussian) inside the channel is considerably larger than the particle. These assumptions are indeed valid in the analyzed case. Because a flow channel with square cross section was used and the dimension of the laser beam was smaller than the width of the channel (250 μm), the changes in dimension and intensity (<4%) of the laser beam due to refraction at the surface of the channel are negligible (unlike the case in cylindrical channels). The bacterial cells were modeled as homogeneous particles using the DDA method (20–22).

The DDA method was first formulated by Purcell and Pennypacker (23), who used it to study interstellar dust grains, and later extended by other researchers such as Draine, and Taubenblatt and Tran (24,25). In DDA an arbitrarily shaped particle is treated as a three-dimensional assembly of dipoles ($j = 1, \ldots, N$) on a cubic grid, located at positions $\mathbf{r}_j$ (26). Each dipole is assigned a complex polarizability α_I, which can be computed from the complex refractive index of the bulk material and the number of dipoles in a unit volume. The dipole moment or polarization at each dipole is related to the electric field by $\mathbf{P}_j = \alpha_j \mathbf{E}_{\text{tot},j}$, where $\mathbf{P}_j$ is the dipole moment at the dipole j, and $\mathbf{E}_{\text{tot},j}$ is the total electric field at dipole j, at $\mathbf{r}_j$.

Following the notation of Ref. (27), the field $\mathbf{E}_{\text{tot},j}$ at each dipole can be decomposed into the electrical field incident upon the features and the electric field contribution from the other interacting dipoles. Hence, the electric field can then be represented as $\mathbf{E}_{\text{tot},j} = \mathbf{E}_{\text{inc},j} + \mathbf{E}_{\text{dipole},j}$, where $\mathbf{E}_{\text{dipole},j}$ is the electric field contribution from the other N–1 dipoles, and $\mathbf{E}_{\text{inc},j}$ is the known incident field $\mathbf{E}_0 \exp(i\mathbf{k}\cdot\mathbf{r}_i - i\omega t)$. Therefore, $\mathbf{E}_{\text{dipole},j}$ can be expressed as:

$$\mathbf{E}_{\text{dipole},j} = \mathbf{E}_{\text{inc},j} - \sum_{k \neq j} \mathbf{A}_{jk}\mathbf{P}_k, \qquad (1)$$

where $\mathbf{A}_{jk}\mathbf{P}_k$ is the electric field at $\mathbf{r}_j$ due to dipole $\mathbf{P}_k$. Each element $\mathbf{A}_{jk}$ is a 3 × 3 matrix:

$$\mathbf{A}_{jk} = \frac{\exp(ikr_{jk})}{r_{jk}} \times \left[k^2(\hat{r}_{jk}\hat{r}_{jk} - \mathbf{1}_3) + \frac{ikr_{jk} - 1}{r_{jk}^2}(3\hat{r}_{jk}\hat{r}_{jk} - \mathbf{1}_3) \right], \quad j \neq k, \qquad (2)$$

where $k \equiv \omega/c$, $r_{jk} = |\mathbf{r}_j - \mathbf{r}_k|$, $\hat{r}_{jk} \equiv (\mathbf{r}_j - \mathbf{r}_k)/r_{jk}$, and $\mathbf{1}_3$ is a 3 × 3 identity matrix. With $\mathbf{A}_{jj} \equiv \alpha_j^{-1}$, the scattering problem is reduced to finding polarizations $\mathbf{P}_j$ that satisfy a system of equations:

$$\sum_{k=1}^{N} \mathbf{A}_{jk}\mathbf{P}_k = \mathbf{E}_{\text{inc},j}. \qquad (3)$$

These equations can be solved by iterations. By introducing the Green function, the method produces reliable results for extremely rough discretization grids such as 2.22 meshes per wavelength (26). In the presented study the quasi-minimal residual method has been used to solve the problem. Owing to the characteristics of the coefficient matrix, the convergence towards an accurate answer is dependent on scattering feature size and refractive index.

Light-scatter signal from four different bacteria species was modeled: *E. coli*, *L. innocua*, *B. subtilis*, and *E. faecalis*. *E. coli*, *L. innocua*, and *B. subtilis* are rod-shaped bacteria, whereas *E. faecalis* appears as cocci in chains. The size of *E. coli* depends on the growth phase, and the nutrients available in the medium. *E. coli* bacilli can be up to 1.5 μm wide and 2.0–6.0 μm long (28). For the purpose of modeling we assumed that *E. coli* cells are typically 2 μm in length and about 1 μm in diameter. *L. innocua* cells were modeled as rods, ∼2 μm in length, and ∼0.6 μm in width (29). *B. subtilis* forms long rods with oval endospores. The dimensions for our model were based on direct observation under a phase-contrast light microscope (Leica Microsystems, Bannockburn, IL): the average size of the vegetative cell was 4.3 μm × 0.54 μm and the endospores measured about 0.8 μm × 0.5 μm. Typically, the volume of the *B. subtilis* cells increases by as much as 4% when spores are formed, whereas the refractive index decreases from 1.51 to 1.39 (30). *E. faecalis* forms oval cocci elongated in the direction of the chain, mostly in pairs and short chains, with each coccus measuring about 1.38 μm long. The refractive indices of vegetative cells of all these bacteria vary from 1.4 to 1.5 (30,31).

The obtained results are valid for forward angles. The effects of internal nonhomogeneity that affect the light scatter at large angles can be ignored in this study. Because the cells are much smaller than the incident laser beam, an incident uniform plane wave is assumed. The numerical model described in this report assumes a nominal effective refractive index of 1.394. The angular variation of scatter has been corrected for refraction of the scattered partial waves across the flow cell on the way to the detectors, and the longer axes of the bacteria were assumed to be aligned with the axis of flow owing to the hydrodynamic forces in a flow cytometer. Because polarization changes the scattering cross section

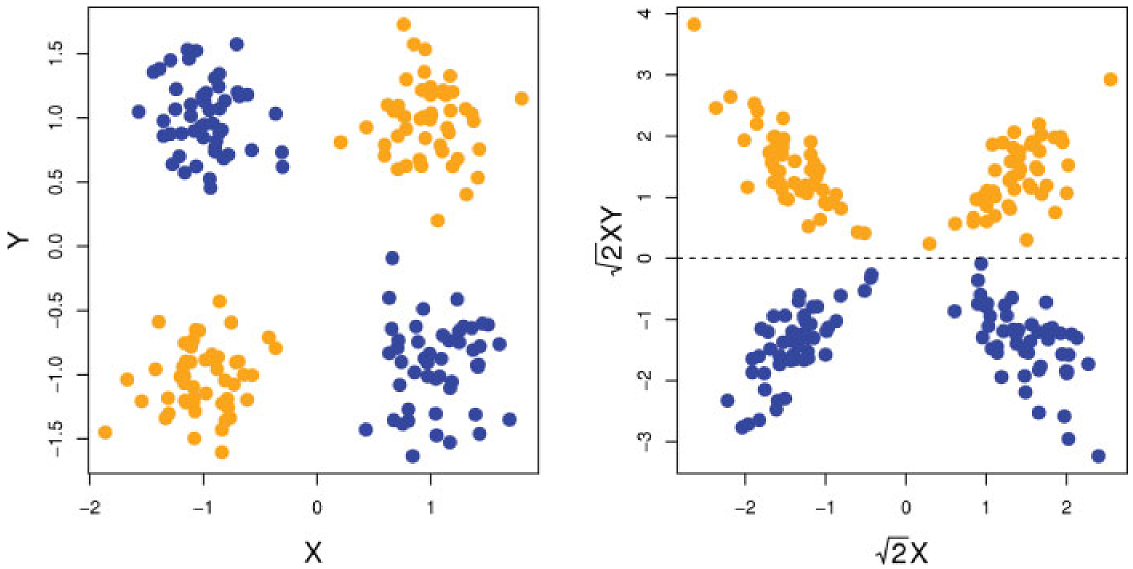

Figure 2. The toy XOR problem demonstrates the concept of mapping the features to higher dimensionality to find a linear separation. The red points (class 1) and green points (class 2) cannot be separated by a linear function in the feature space (left plot). However, a simple mapping to a higher dimension allows linear separation (right plot). The classes can be mapped to a six-dimensional space: 1, $\sqrt{2}$X, $\sqrt{2}$Y, $\sqrt{2}$XY, X2, Y2, where the optimal separation hyperplane is $XY = 0$. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

noticeably, especially for long rod-like particles, the employed model takes into account incident laser beam polarization ($\mathbf{E}_x/\mathbf{E}_y = 0.33$).

Machine Learning Tools

Among various machine learning tools tested for classification of scatter features of individual bacteria, SVM-based algorithms were especially promising (32–34). SVM algorithms allow for nonlinear decision boundaries in the input space. SVMs are based on the concept of decision hyperplanes that define decision boundaries. A decision hyperplane is one that separates a set of objects having different class memberships. SVMs are able to construct hyperplanes in a multidimensional space that separates cases of different class labels. An optimal decision hyperplane is here defined as the linear decision function with maximal margin between the vectors of the two classes. It has been demonstrated that to construct such hyperplanes one has to take into account only a small amount of the training data, the so-called support vectors, which determine this margin (33). For $\mathbf{w}_0 \cdot \mathbf{z} + b_0 = 0 \mid \mathbf{w} \in R^N$, $b \in R$, which is the optimal hyperplane, it has been shown that the weights $\mathbf{w}_0$ can be expressed as linear combination of support vectors:

$$\mathbf{w}_0 = \sum_{\text{support vectors}} \alpha_i \mathbf{z}_i \tag{4}$$

Therefore, the linear decision function $I(\mathbf{z})$ will be in the form of

$$I(\mathbf{z}) = \operatorname{sign}\left(\sum_{\text{support vectors}} \alpha_i \mathbf{z}_i \cdot \mathbf{z} + b_0\right) \tag{5}$$

where $\mathbf{z}_i \cdot \mathbf{z}$ is the dot-product between support vectors $\mathbf{z}_i$ and vector $\mathbf{z}$ in feature space. SVM is a linear classifier in the parameter space, but it is easily extended to a nonlinear classifier by mapping the space $S = \{\mathbf{x}\}$ of the input data into a high-dimensional (possibly infinite-dimensional) feature space $F = \{\phi(\mathbf{x})\}$ (see Fig. 2). If one chooses an adequate mapping ϕ, the data points become linearly separable or mostly linearly separable in the high-dimensional space, so that one can easily apply the structure risk minimization (35). To avoid working in the potentially high-dimensional space F, one tries to pick a feature space in which the dot product can be evaluated directly using a nonlinear function in input space, i.e. by means of the kernel trick: $\kappa(\mathbf{x}_1, \mathbf{x}_2) = \langle\phi(\mathbf{x}_1), \phi(\mathbf{x}_2)\rangle$. Therefore, instead of making a nonlinear transformation of the input vectors followed by dot-products with support vectors in feature space, one can first compare two vectors in input space, and then make a nonlinear transformation of the value of the result (33). A kernel can be also understood as a similarity measure between two observations. A large value for $\kappa(\mathbf{x}_1,\mathbf{x}_2)$ indicates similar points, where smaller values indicate dissimilar points. Typical kernels include the linear kernel, $\kappa(\mathbf{x}_1, \mathbf{x}_2) = \mathbf{x}_1{}^{\mathrm{T}} \mathbf{x}_2$, the polynomial kernel, $\kappa(\mathbf{x}_1,\mathbf{x}_2) = (\mathbf{x}_1{}^{\mathrm{T}} \mathbf{x}_2 + 1)^{\mathrm{d}}$, or the RBF kernel, $\kappa(\mathbf{x}_1, \mathbf{x}_2) = \exp(-\gamma\|\mathbf{x}_1 - \mathbf{x}_2\|^2)$. It has been shown that all these kernels are functions of dot products (36).

Supervised classification performed in this report used an implementation SVM-based algorithm by Chih-Chung Chang and Chih-Jen Lin (37–39) All the plots, including the exam-

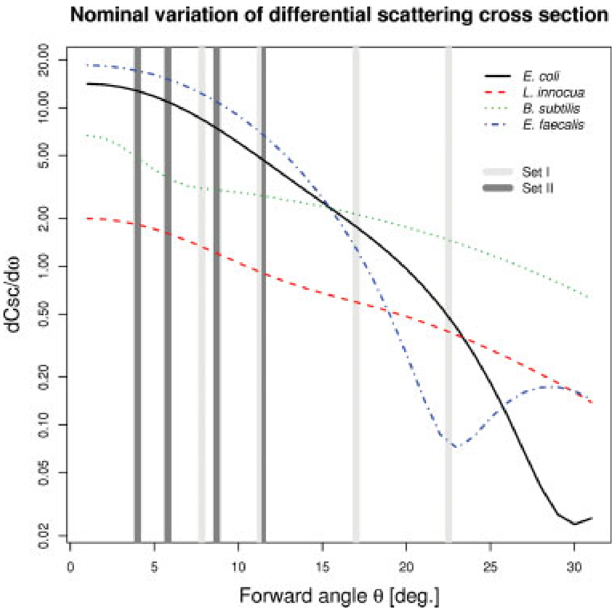

Figure 3. The nominal variation of differential scattering cross section ($dCsc/d\omega$) with forward angle for the four bacteria species. Set I—7.8, 11.3, 17, and 22.5°; Set II—4, 5.8, 8.7, and 11.5°. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

ples of the SVM decision boundaries, were prepared using R, a free software environment for statistical computing and graphics (40).

Results

Calculation of Distinguishability Factor

Light-scatter patterns created by cells belonging to four different bacterial species, *L. innocua*, *B. subtilis*, *E. coli*, and *E. faecalis*, were modeled using the DDA method. Flow cytometry measurements of traditional forward- and side-scatter signals, as well as multiangle scatter measurements, have been performed using actual samples of *E. coli* K12, *L. innocua* F4248, *B. subtilis* ATCC 6633, and *E. faecalis* CG110.

The employed model of scattering by bacteria with a nominal refractive index predicts scatter intensities at angles varying from 0 to 30°. The predictions of angular scatter intensity are shown in Figure 3 (averaged over all Φ for each ring). The plotted differential scattering cross section is independent of the distance between the detector and the sample, and is used to select the angles for maximum discrimination of the bacteria.

The analytical model of the bacteria with a nominal refractive index and size allowed us to find the optimal location for the scatter detectors for effective classification. The proper placement of the detectors was determined by the value of distinguishability factor *D*, defined as the ratio of the difference in scattering cross section to the sum of scattering cross section of two different bacteria:

$$D = \sum_{\theta} \left[(dCsc_i - dCsc_j)/(dCsc_i + dCsc_j) \right], \qquad (6)$$

where *i,j* represent different bacterial species, and θ is the angle of light scatter.

The idea of calculating the *D* factor can be easily derived from an analysis of Figure 3. One can easily see that the angles represented by light-gray lines (nominal 7.8, 11.3, 17, and 22.5°—Set I) are better than the angles highlighted by dark-gray lines (4, 5.8, 8.7, and 11.5°—Set II) for distinguishing all the analyzed two-component mixtures of bacteria. These sets of angles can in practice be translated to different positions of the multi-angle scatter detector. The distinguishabilities for the two analyzed sets are presented in Table 1.

The data showed that the ability to distinguish between the analyzed bacterial species decreased when the multiangle detector was placed too close to the flow chamber, effectively collecting signals from larger angles of scatter. *L. innocua* was an exception from this finding.

Although this estimate ignores the variation of scatter signals within each population due to intra-species differences in size and refractive indices, it gave a qualitative approximation of the expected outcome of the alternative measurement scenarios.

The difference in refractive indices and cell sizes results in dispersion of each bacterial species population in the light-scatter measurement space. Therefore, the experimental data have not been directly classified using the output of a model, but rather processed employing a machine learning system.

Predicted Classification Success

To predict the feasible classification, the variability in scatter signal owing to differences in refractive index and sizes of individual particles had to be considered. Hence, a normal distribution of sizes and refractive indices was employed in the enhanced model calculated for three species of bacteria (*L. innocua*, *E. faecalis*, and *E. coli*). The 1/*e* width of the normal distribution of refractive index was 0.033 with a mean (μ) of 1.394. The standard deviation of the refractive index was assumed to be approximately 2%. Similarly, the standard

Table 1. Distinguishabilities calculated for scatter measurement at nominal 7.8, 11.3, 17, and 22.5° (Set I), and 4, 5.8, 8.7, and 11.5° (Set II)

	E. COLI		*L. INNOCUA*		*B. SUBTILIS*		*E. FAECALIS*	
	I	II	I	II	I	II	I	II
E. coli			1.88	2.82	2.15	1.49	1.34	0.67
L. innocua	1.88	2.82			2.11	1.79	2.42	3.12
B. subtilis	2.15	1.49	2.11	1.79			2.25	2.01
E. faecalis	1.34	0.67	2.42	3.12	2.25	2.01		

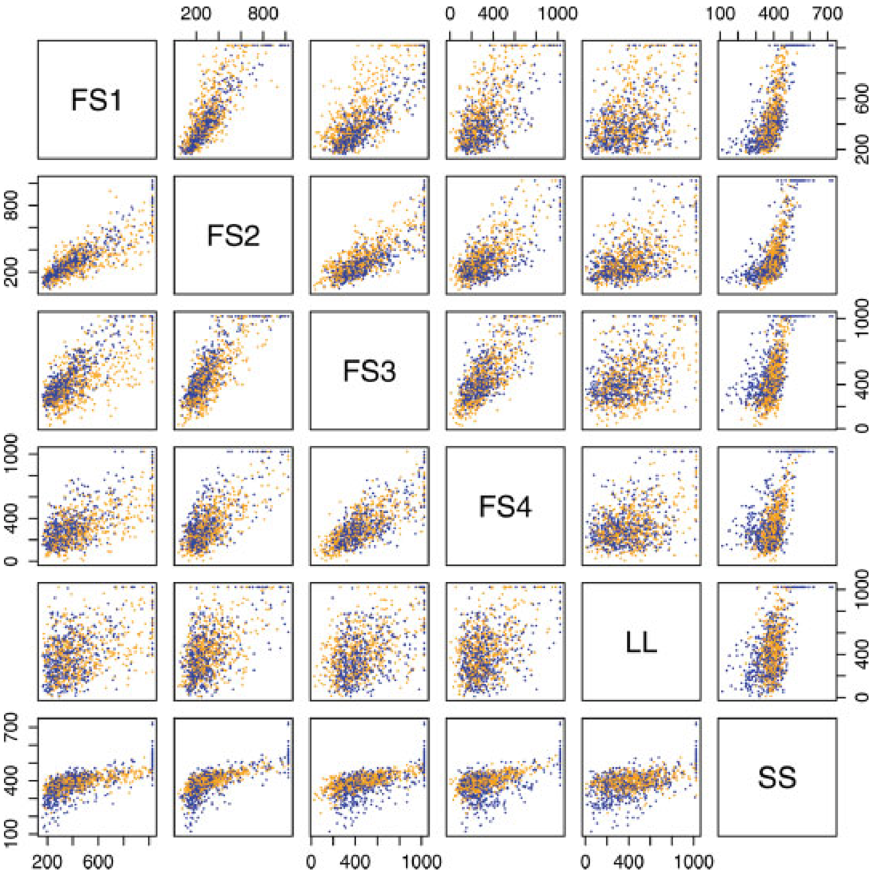

Figure 4. Matrix of scatter plots representing multiangle scatter measurement of *B. subtilis* (red dots) and *E. faecalis* samples (green dots). For clarity, only 1,000 events were plotted. FS14, four forward-scatter measurements; LL, axial light loss; SS, side scatter. *B. subtilis*—orange dots, *E. faecalis*—blue dots. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

deviation in volume of the bacteria was assumed to be 5%. This increase in volume was modeled by corresponding isotropic changes in the dimensions of the bacteria. The modeled populations of each bacterial species were divided into 91 subgroups (13 different values of refractive index by 7 values volume, both varying from $\mu - 3\sigma$ to $\mu + 3\sigma$) and mapped onto the four-dimensional measurement space of the two investigated angle sets. The modeled bacterial populations were used to calculate theoretical scatter signals using the DDA approach as described before. Once scatter intensities of each subgroup were computed, the results were weighted by the population density (w) using the equation

$$w_{ij} = \int_{\text{subgroup}} \frac{1}{\sigma_n\sqrt{2\pi}} \exp\left(-\frac{(n_i - n_0)^2}{2\sigma_n^2}\right) \frac{1}{\sigma_V\sqrt{2\pi}} \times \exp\left(-\frac{(V_i - V_0)^2}{2\sigma_V^2}\right), \quad (7)$$

where n_i and V_i are the refractive index and volume of the ij-th subgroup, n_0 and σ_i are the mean and the standard deviation of refractive index, V_0 and σ_V are the mean and the standard deviation of bacterial cell volume.

The overlap of weighted resultant values calculated for every pair of bacterial species in the measurement space was used as an estimate of the possible classification error, for every two-class case. Because the model did not take the instrument noise into account, these values were expected to give an approximate upper bound of the feasible classification success.

In Silico Analysis of Flow Cytometry Experiments

Samples containing the pure bacterial suspensions were run in sequence but separately on the modified Beckman-Coulter FC500 flow cytometer. Subsequently, the datasets were electronically mixed, and a parameter representing the ground truth was added to the dataset. This parameter was used to verify the results of automated classification.

The collected scatter signals at the four forward angles established by the numerical analysis study, a parameter representing sum of all the forward-scatter intensities, side-scatter, and axial light loss measures for each particle from every group of bacteria formed multidimensional data vectors describing the analyzed bioparticles. Visual examination of plots representing measurements of forward- and side-scatter signals could not distinguish between the microbial particles of *E. coli* K12, *L. innocua* F4248, *B. subtilis* ATCC 6633, and

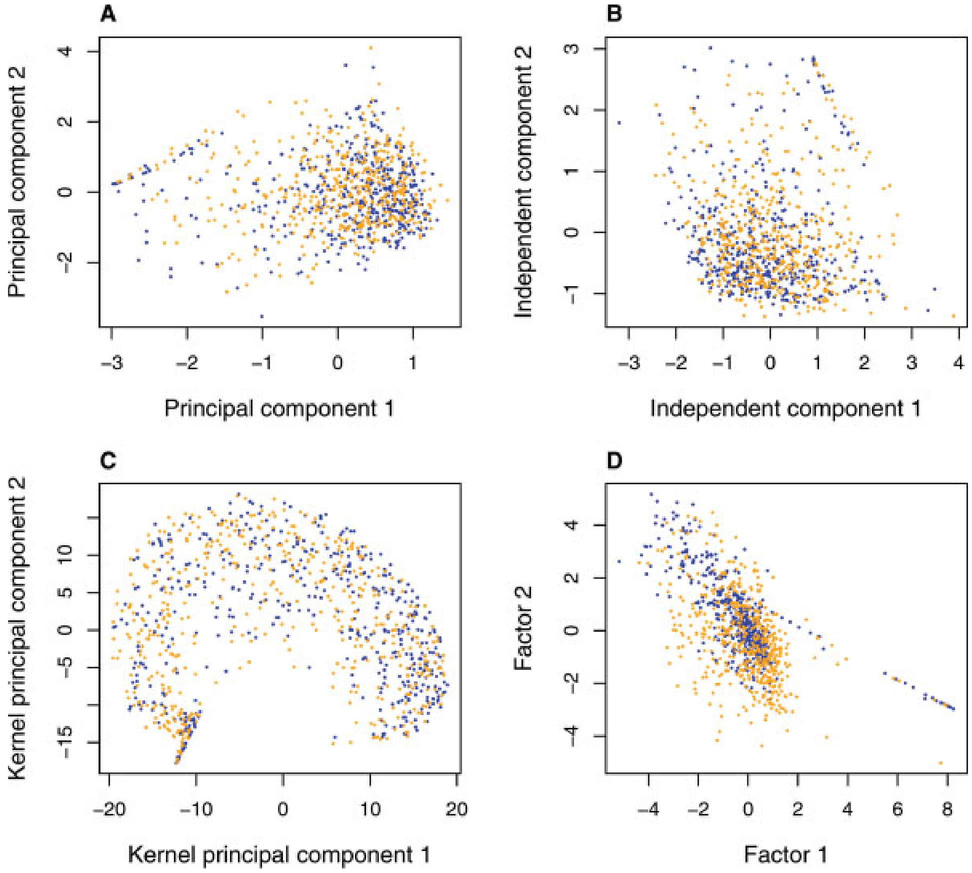

Figure 5. Example of dimensionality reduction techniques applied to the multiangle scatter data. (**A**) Principal component analysis, (**B**), independent component analysis, (**C**), kernelized version of principal component analysis, (**D**), factor analysis. *B. subtilis*—orange dots, *E. faecalis*—blue dots. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

E. faecalis CG110 in any of the experiments. An example is demonstrated in Figure 4, where samples containing particles of *E. faecalis* and *B. subtilis* are represented on a scatter-plot matrix. Partial or complete overlap of the two populations in the parametric space is evident.

The classification problem was then to determine the type (species) of every analyzed particle on the basis of its multidimensional data vectors. The unsupervised dimensionality reduction approach employing linear and kernel principal component analysis using radial basis function (PCA, kPCA), independent component analysis (ICA), as well as factor analysis (FA) have not resulted in separable populations (see an example in Fig. 5). Attempted supervised classification using linear discriminant analysis (LDA) also failed, producing results with error rates above 35% (Fig. 6) in all the cases tested.

In contrast to LDA, SVM is usually capable of solving complex classification problems which do not have simple linear (or quadratic) solution in the parametric space (Fig. 2). Therefore, supervised classification was performed using an SVM-based approach. A radial-basis function kernel $\kappa(\mathbf{x}_1,\mathbf{x}_2) = \exp(-\gamma\|\mathbf{x}_1 - \mathbf{x}_2\|^2)$ was used for all the classification. The optimal type of kernel was established experimentally. The SVM complexity parameter as well as the γ kernel parameter was found by an extensive grid search evaluating every pair of parameters by re-training and cross-validation.

The 5 × 2 cross-validation and bootstrap algorithms were used to determine the classification success of the optimal SVM. The accuracy of classification computed using cross-validation is summarized in Table 2. An example of a complicated decision boundary (a hyperplane in n dimensions, where n is the number of parameters) determined by a typical SVM training applied to a scattered-light dataset is illustrated in Figure 7.

Discussion and Future Research

Although light-scatter signatures of cells have been utilized in microbiological applications of flow cytometry, the role of scattered light was secondary at best. It was the growing availability of fluorescence-labeled antibodies to specific antigens that made possible the use of flow cytometry to directly detect the presence of pathogens. Cytometry-based methods have been employed to detect surface antigens in *Haemophilus* (41), *Salmonella* (42,43), *Mycobacterium* (44), *Brucella* (45), *Branhamella catarrhalis* (46), *Mycoplasma fermentans* (47), *Pseudomonas aeruginosa* (48), *Bacteroides fragilis* (49,50), *Legionella* (51), and other microorganisms (52). The main dis-

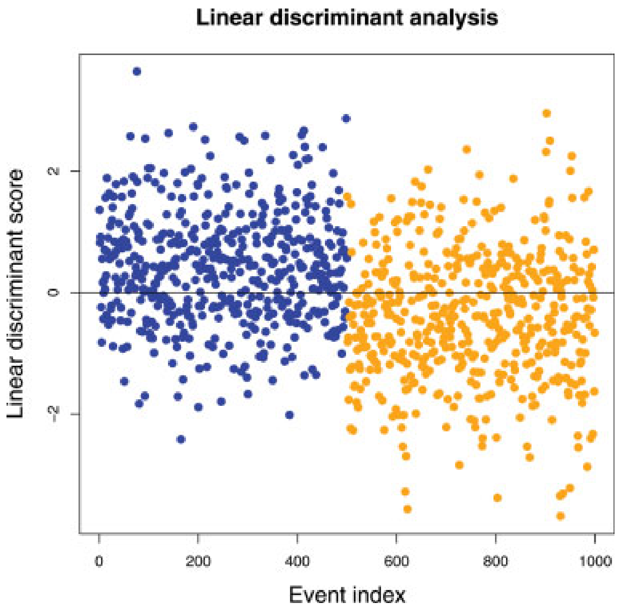

Figure 6. Linear discriminant score plot illustrating inability to achieve separation between data vectors describing *B. subtilis* (orange dots) and *E. faecalis* (blue dots). Events indexed from 1 to 500 should be placed above the $y = 0$ function, whereas all the events from 500 to 1000 should score below 0. However, owing to misclassification a large portion of dots representing *B. subtilis* is placed above $y = 0$ discriminant function. Conversely, a large group of *E. faecalis* cells was misclassified as *B. subtilis*. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

advantages of label-dependent detection are the limited availability of antibodies directed against certain microorganisms, and problems with fluorescence detection multiplexing. Although scatter signals have been routinely collected during flow cytometry measurements of bacterial populations, classification of live microorganisms on the basis of scatter signal alone measured in a commercially available flow system has not been reported.

Since the pioneering experiments by Salzmann et al., (9) it has been demonstrated in numerous reports that label-free measurements and classification of bioparticles in flow cytometry is feasible. The major obstacles for the wider implementation of the multiangle scatter systems were the complexity of the design and the lack of easy-to-use tools for data analysis. Although the system reported in this manuscript uses an enhanced detector, the number of simultaneously measured angles is relatively low, and detector installation does not require extensive modifications to the flow cytometry hardware. Instead of focusing on the increase of the number of scatter angles the proposed approach requires pre-selecting angles which are likely to offer high distinguishability for the bioparticles of interest. This is a strength of our design, but also a weakness, since the system can be optimally set up only for a given (and known) type of bioparticle. Consequently, the system as proposed cannot be used for purely exploratory flow cytometry, in which the characteristics of the analyzed bioparticles are completely unknown. However, if control samples are available, and particles whose presence has to be determined (or which have to be enumerated) can be characterized in terms of their scatter properties, there is a good chance that a system can be tuned to accommodate such a specialized measurement. Alternatively, one may locate the optimal position of the detector (and consequently, the collected scatter angles), simply by trial and error, where results obtained from controls are electronically mixed, and classified with a machine-learning system, using cross-validation to determine the optimal angles. The current prototype used for demonstration of the proof of concept allows for only two positions of the detectors, but there is absolutely no technical reason why multiple positions along the *z*-axis and consequently multiple sets of angles could not available.

Comparison of the classification success obtained experimentally to the distinguishabilities estimated from the simple scatter model shows high level of agreement except for two of the six classification cases for each set of angles. This is encouraging considering the fact that the intra-population variance in size and refractive indices has not been accounted for in the first model. It should also be noted that the predicted high classification rates for *L. innocua* and *E. faecalis* mixture measured with the second configuration (4°, 5.8°, 8.7°, and 11.5°), do not match the experimental results (Table 3). We suspect that the reason for this discrepancy is the high variance in dimensions and refractive index of bacteria.

The upper bound of classification success estimated with the help of the enhanced model for turned out to be valid, although over-optimistic. The real classification success differed from the estimated by 1–10%. However, we still consider the model to have high predictive power since only in one of the analyzed cases (*B. subtilis* vs. *L. innocua*) was the predicted classification rate lower than the real accuracy (Table 2). This

Table 2. Average classification success rates for 6-parameter (7.8, 11.3, 17.7, 22.5, 90°, and axial light loss) scatter system employing SVM classifier

	E. COLI		*L. INNOCUA*		*B. SUBTILIS*		*E. FAECALIS*	
	R (%)	*E* (%)	*R* (%)	*E* (%)	*R* (%)	*E* (%)	*R* (%)	*E* (%)
E. coli	–	–	86.30	95.8	99.10	100	68.70	77.1
L. innocua	86.30	95.8	–	–	99.60	98	81.60	95.6
B. subtilis	99.10	100	99.60	98	–	–	98.50	100
E. faecalis	68.70	77.1	81.60	95.6	98.50	100	–	–

R, real (measured) classification accuracy; *E*, estimated classification accuracy.

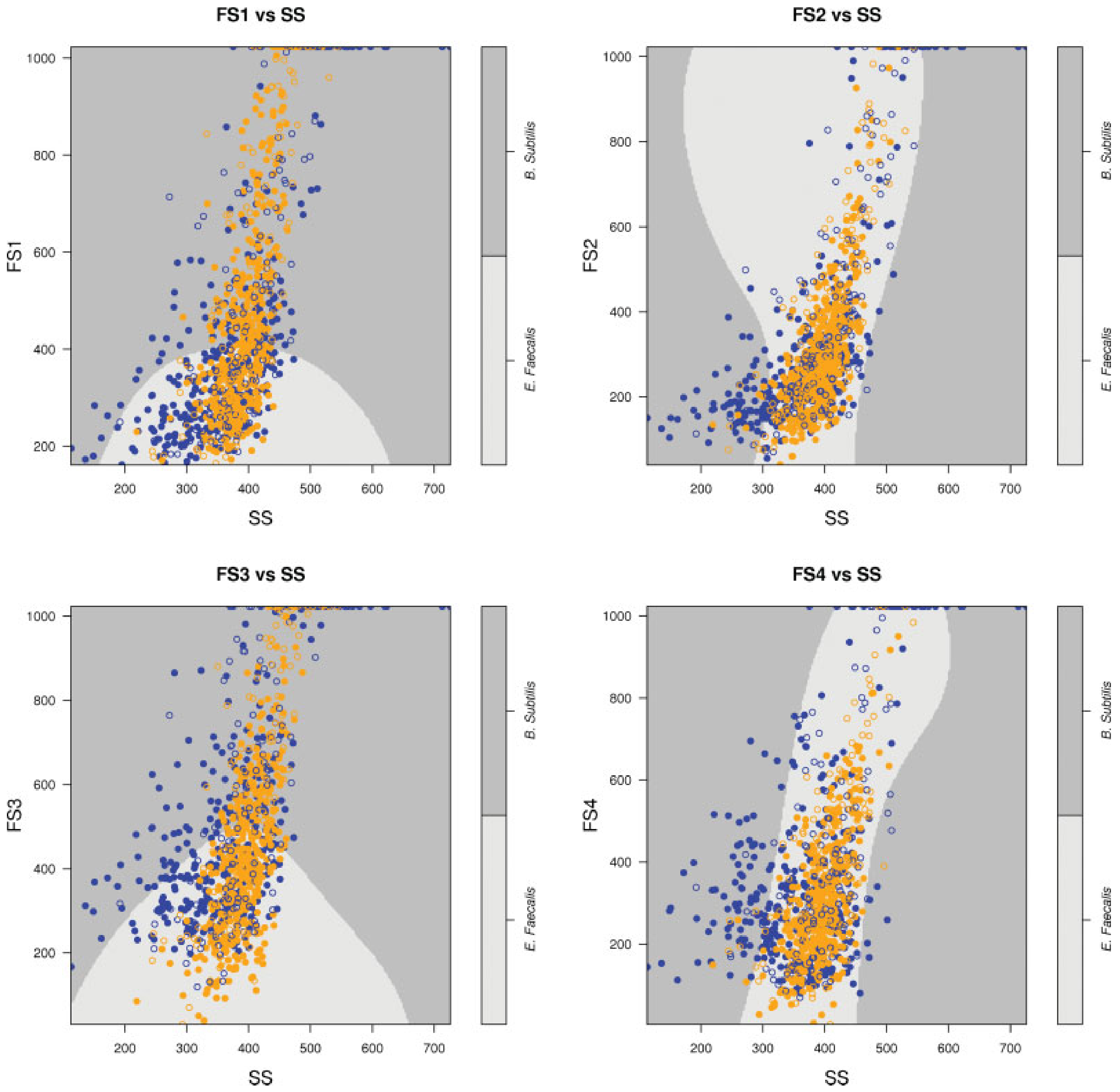

Figure 7. Examples of cross sections through decision boundaries of SVM-based pattern-recognition system. Filled points represent regular data vectors, empty points represent support vectors. Values of the variables not represented on the 2D plots were set to their medians. *B. subtilis*—orange points, *E. faecalis*—blue points. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

shows that the simulated upper bound can be used to determine a priori whether a certain type of analysis and classification is feasible in the given system. For instance, if the simulated upper bound is on the level of 65–70%, any attempt for successful classification will be most likely futile regardless of the quality of the sample and stability of the lasers.

In the presented report the scatter simulation employing state-of-the art techniques such as DDA allowed us to utilize optimally the multiangle detector. However, owing to high biological variability of the real samples containing microorganisms we have not employed scatter simulation for the purpose of the actual particle classification. Instead, a machine-learning system was used.

Machine-learning and pattern recognition systems have been applied to flow cytometry by a number of researchers in fields such as marine biology (53,54), hematology and immunology (55,56), and microbiology (57,58). Among the proposed methods were LDA, neural networks, and SVM (59). However, we are not aware of any application of these techniques to label-free bacteria classification. Another aspect of this

Table 3. Average classification success rates for 6-parameter (4, 5.8, 8.7, 11.5°, and axial light loss) scatter system employing SVM classifier

	E. COLI		*L. INNOCUA*		*B. SUBTILIS*		*E. FAECALIS*	
	R (%)	*E* (%)	*R* (%)	*E* (%)	*R* (%)	*E* (%)	*R* (%)	*E* (%)
E. coli	–	–	74.8	79.0	69.9	81.0	57.9	61.0
L. innocua	74.8	79.0	–	–	71.7	74.0	77.0	79.0
B. subtilis	69.9	81.0	71.7	74.0	–	–	70.2	87.0
E. faecalis	57.9	61.0	77.0	79.0	70.2	87.0	–	–

R, real (measured) classification accuracy; *E*, estimated classification accuracy.

work is a combination of simulation-based pre-selection of features with a machine-learning system. The premise of this approach is two-fold. Firstly, the smaller number of parameters to collect simplifies the design of the detector, making it pluggable to older hardware. Secondly, overwhelming a machine-learning system with nonrelevant features may degrade the performance of classifiers. Naturally, employing a feature selection procedure is an answer to the problem of a huge number of features. However, this solution comes at significant computational cost, and ultimately it may be difficult to implement, especially if a real-time analysis or fast analysis and classification is desirable.

Use of SVM allowed for high classification accuracy and eliminated the need for gating. Manual gating can be performed easily if analyzed parameters are orthogonal. However, intensity of light collected by scatter detectors cannot be orthogonalized via compensation as in the case of fluorescence measurements. Therefore, other methods of classification had to be explored. PCA, kPCA, ICA, and FA failed to separate the analyzed populations. A simple linear discrimination approach was also unable to perform in a satisfactory manner. However, supervised classification employing a kernel approach, such as SVM, produced a remarkably high success rate (Table 2). Unfortunately, SVM results cannot be easily interpreted if the dimensionality of the problem is higher than 2 (compare Fig. 2 with Fig. 7). This may be a serious problem for many practitioners in the field who expect that despite the growth in the number of available variables, some simple graphical model of data analysis would still be employed. Therefore, one of the most important aspects of multiangle scatter studies should be a search for innovative data visualization tools, allowing for meaningful dimensionality reduction and easy exploratory gating.

Literature Cited

1. Salzman GC, Crowell JM, Martin JC, Labauve PM, Mullaney PF. Classification of human leukocytes by multiangle laser light-scattering in a flow system. Biophys J 1975;15:A240.
2. Visser JWM, Engh GJVD, Bekkum DWV. Light-scattering properties of murine hematopoietic cells. Blood Cells 1980;6:391–407.
3. Dubelaar GBJ, Visser JWM, Donze M. Anomalous behavior of forward and perpendicular light-scattering of a cyanobacterium owing to intracellular gas vacuoles. Cytometry 1987;8:405–412.
4. Mullaney PF, Dean PN. Cell sizing—A small-angle light-scattering method for sizing particles of low relative refractive index. Appl Opt 1969;8:2361–2362.
5. Mullaney PF, Vandilla MA, Coulter JR, Dean PN. Cell sizing—A light scattering photometer for rapid volume determination. Rev Sci Instrum 1969;40:1029–1032.
6. Kerker M, Chew H, Mcnulty PJ, Kratohvil JP, Cooke DD, Sculley M, Lee MP. Light-scattering and fluorescence by small particles having internal structure. J Histochem Cytochem 1979;27:250–263.
7. Kamentsky LA, Melamed MR, Derman H. Spectrophotometer—New Instrument for ultra rapid cell analysis. Science 1965;150:630–631.
8. Meyer RA, Haase SF, Poduslo SE, Mckhann GM. Light-scattering patterns of isolated oligodendroglia. J Histochem Cytochem 1974;22:594–597.
9. Salzman GC, Crowell JM, Mullaney PF. Flow-system multi-angle light-scattering instrument for biological cell characterization. J Opt Soc Am 1975;65:1170–1171.
10. Loken MR, Sweet RG, Herzenberg LA. Cell discrimination by multiangle light-scattering. J Histochem Cytochem 1976;24:284–291.
11. Bartholdi M, Salzman GC, Hiebert RD, Kerker M. Differential light-scattering photometer for rapid analysis of single particles in flow. Appl Opt 1980;19:1573–1581.
12. Salzman GC, Crowell JM, Goad CA, Hansen KM, Hiebert RD, Labauve PM, Martin JC, Ingram ML, Mullaney PF. Flow-system multiangle light-scattering instrument for cell characterization. Clin Chem 1975;21:1297–1304.
13. Salzman GC, Crowell JM, Hansen KM, Ingram M, Mullaney PF. Gynecologic specimen analysis by multiangle light-scattering in a flow system. J Histochem Cytochem 1976;24:308–314.
14. Salzman GC, Burger DE, Bartholdi M. Light-scattering from single particles and biological cells in a flow system. J Opt Soc Am 1977;67:1382.
15. Shvalov AN, Surovtsev IV, Chernyshev AV, Soini JT, Maltsev VP. Particle classification from light scattering with the scanning flow cytometer. Cytometry 1999;37:215–220.
16. Shvalov AN, Soini JT, Surovtsev IV, Kochneva GV, Sivolobova GF, Petrov AK, Maltsev VP. Individual *Escherichia coli* cells studied from light scattering with the scanning flow cytometer. Cytometry 2000;41:41–45.
17. Yurkin MA, Semyanov KA, Tarasov PA, Chernyshev AV, Hoekstra AG, Maltsev VP. Experimental and theoretical study of light scattering by individual mature red blood cells by use of scanning flow cytometry and a discrete dipole approximation. Appl Opt 2005;44:5249–5256.
18. Steen HB. Flow cytometer for measurement of the light scattering of viral and other submicroscopic particles. Cytometry Part A 2004;57A:94–99.
19. Maltsev VP. Scanning flow cytometry for individual particle analysis. Rev Sci Instrum 2000;71:243–255.
20. Schmehl R, Nebeker BM, Hirleman ED. Discrete-dipole approximation for scattering by features on surfaces by means of a two-dimensional fast Fourier transform technique. J Opt Soc Am A: Opt Image Sci Vision 1997;14:3026–3036.
21. Nebeker BM, Starr GW, Hirleman ED. Evaluation of iteration methods used when modeling scattering from features on surfaces using the discrete-dipole approximation. J Quant Spectrosc Radiat Transfer 1998;60:493–500.
22. Nebeker BM, de la Pena JL, Hirleman ED. Comparisons of the discrete-dipole approximation and modified double interaction model methods to predict light scattering from small features on surfaces. J Quant Spectrosc Radiat Transfer 2001; 70:749–759.
23. Purcell EM, Pennypacker CR. Scattering and absorption of light by nonspherical dielectric grains. Astrophys J 1973;186:705–714.
24. Draine BT. The discrete-dipole approximation and its application to interstellar graphite grains. Astrophys J 1988;333:848–872.
25. Taubenblatt MA, Tran TK. Calculation of light-scattering from particles and structures on a surface by the coupled-dipole method. J Opt Soc Am A: Opt Image Sci Vision 1993;10:912–919.
26. Wriedt T. A review of elastic light scattering theories. Part Part Syst Characterization 1998;15:67–74.
27. Draine BT, Flatau PJ. Discrete-dipole approximation for scattering calculations. J Opt Soc Am A: Opt Image Sci Vision 1994;11:1491–1499.
28. Van De Merwe WP, Czege J, Milham ME, Bronk BV. Rapid optically based measurements of diameter and length for spherical or rod-shaped bacteria in vivo. Appl Opt 2004;43:5295–5302.
29. Gupta A, Akin D, Bashir R. Detection of bacterial cells and antibodies using surface micromachined thin silicon cantilever resonators. J Vacuum Sci Technol B 2004;22:2785–2791.
30. Katz A, Alimova A, Xu M, Gottlieb P, Rudolph E, Steiner JC, Alfano RR. In situ determination of refractive index and size of Bacillus spores by light transmission. Opt Lett 2005;30:589–591.
31. Wyatt PJ. Differential light scattering—A physical method for identifying living bacterial cells. Appl Opt 1968;7:1879–1895.

32. Vapnik V. The Nature of Statistical Learning Theory, 2nd ed. New York: Springer; 2000.
33. Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995;20:273–297.
34. Duda RO, Hart PE, Stork DG. Pattern Classification, 2nd ed. New York: Wiley; 2001.
35. Amari S, Wu S. Improving support vector machine classifiers by modifying kernel functions. Neural Net 1999;12:783–789.
36. Scholkopf B, Smola AJ. A short introduction to learning with kernels. Adv Lect Mach Learn 2002;2600:41–64.
37. Chang CC, Lin CJ. LIBSVM: A library for support vector machines. http://www.csie.ntu.edu.tw/~cjlin/libsvm/; 2001.
38. Burges ChJC. A tutorial on support vector machines for pattern recognition. Data Min Knowl Discov 1998;2:121–167.
39. Fan RE, Chen PH, Lin CJ. Working set selection using second order information for training support vector machines. J Mach Learn Res 2005;6:1889–1918.
40. R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org; 2007.
41. Srikumar R, Chin AC, Vachon V, Richardson CD, Ratcliffe MJ, Saarinen L, Kayhty H, Makela PH, Coulton JW. Monoclonal antibodies specific to porin of *Haemophilus influenzae* type b: Localization of their cognate epitopes and tests of their biological activities. Mol Microbiol 1992;6:665–676.
42. Clarke RG, Pinder AC. Improved detection of bacteria by flow cytometry using acombination of antibody and viability markers. J Appl Microbiol 1998;84:577–584.
43. McClelland RG, Pinder AC. Detection of low levels of specific *Salmonella* species by fluorescent antibodies and flow cytometry. J Appl Bacteriol 1994;77:440–447.
44. Ozanne V, Ortalo-Magne A, Vercellone A, Fournie JJ, Daffe M. Cytometric detection of mycobacterial surface antigens: Exposure of mannosyl epitopes and of the arabinan segment of arabinomannans. J Bacteriol 1996;178:7254–7259.
45. Bowden RA, Cloeckaert A, Zygmunt MS, Bernard S, Dubray G. Surface exposure of outer membrane protein and lipopolysaccharide epitopes in *Brucella* species studied by enzyme-linked immunosorbent assay and flow cytometry. Infect Immun 1995; 63:3945–3952.
46. Bhushan R, Kirkham C, Sethi S, Murphy TF. Antigenic characterization and analysis of the human immune response to outer membrane protein E of *Branhamella catarrhalis*. Infect Immun 1997;65:2668–2675.
47. Cheek RF, Olszak I, Madoff S, Preffer FI. In vitro detection of *Mycoplasma fermentans* binding to B-lymphocytes in fresh peripheral blood using flow cytometry. Cytometry 1997;28:90–95.
48. Hughes EE, Matthews-Greer JM, Gilleland HE Jr. Analysis by flow cytometry of surface-exposed epitopes of outer membrane protein F of *Pseudomonas aeruginosa*. Can J Microbiol 1996;42:859–862.
49. Lutton DA, Patrick S, Crockard AD, Stewart LD, Larkin MJ, Dermott E, McNeill TA. Flow cytometric analysis of within-strain variation in polysaccharide expression by *Bacteroides fragilis* by use of murine monoclonal antibodies. J Med Microbiol 1991;35:229–237.
50. Patrick S, Stewart LD, Damani N, Wilson KG, Lutton DA, Larkin MJ, Poxton I, Brown R. Immunological detection of *Bacteroides fragilis* in clinical samples. J Med Microbiol 1995;43:99–109.
51. Ingram M, Cleary TJ, Price BJ, Price RL III, Castro A. Rapid detection of *Legionella pneumophila* by flow cytometry. Cytometry 1982;3:134–137.
52. Alvarez-Barrientos A, Arroyo J, Canton R, Nombela C, Sanchez-Perez M. Applications of flow cytometry to clinical microbiology. Clin Microbiol Rev 2000;13:167–195.
53. Morris CW, Autret A, Boddy L. Support vector machines for identifying organisms—A comparison with strongly partitioned radial basis function networks. Ecol Model 2001;146:57–67.
54. Wilkins MF, Boddy L, Morris CW, Jonker RR. Identification of phytoplankton from flow cytometry data by using radial basis function neural networks. Appl Environ Microbiol 1999;65:4404–4410.
55. Adjouadi M, Zong N, Ayala M. Multidimensional pattern recognition and classification of white blood cells using support vector machines. Part Part Syst Characterization 2005;22:107–118.
56. Toedling J, Rhein P, Ratei R, Karawajew L, Spang R. Automated in-silico detection of cell populations in flow cytometry readouts and its application to leukemia disease monitoring. BMC Bioinform 2006;7:282.
57. Davey HM, Kell DB. Flow cytometry and cell sorting of heterogeneous microbial populations: The importance of single-cell analyses. Microbiol Rev 1996;60:641–696.
58. Davey HM, Jones A, Shaw AD, Kell DB. Variable selection and multivariate methods for the identification of microorganisms by flow cytometry. Cytometry 1999;35:162–168.
59. Boddy L, Wilkins MF, Morris CW. Pattern recognition in flow cytometry. Cytometry 2001;44:195–209.