

Statistical Imaging for Modeling and Identification of Bacterial Types

Sigal Trattner^a, Hayit Greenspan^{a,*}, Gabi Tepper^b, Shimon Abboud^a

^aDepartment of Biomedical Engineering, Faculty of Engineering,
Tel-Aviv University, Tel-Aviv 69978, Israel

^bSpring Diagnostics Ltd., shin ben-zion 51, Rehovot, Israel, 76472

*hayit@eng.tau.ac.il; <http://www.eng.tau.ac.il/~hayit/>

Abstract. An automatic tool is developed to identify microbiological data types using computer-vision and statistical modeling techniques. In bacteriophage (phage) typing, representative profiles of bacterial types are extracted. Currently, systems rely on the subjective reading of the profiles by a human expert. This process is time-consuming and prone to errors. The statistical methodology presented in this work, provides for an automated, objective and robust analysis of the visual data, along with the ability to cope with increasing data volumes. Validation is performed by a comparison to an expert manual segmentation and labeling of the phage profiles.

1 Introduction

The need to analyze vast amounts of microbiological data requires computer modeling and automation. Current manual procedures are prone to large variability within and across the human experts due to natural fuzziness present in the microbiological data. These procedures are time consuming and are of great cost. Reducing the amount of human intervention in the data analysis is crucial in order to cope with the increasing volume of data and to achieve more objective and quantitatively accurate measurements as well as to obtain repeatable results. In this work we combine image analysis with statistical modeling tools in a general framework for visual array analysis. We focus on microbiological data and bacterial type modeling.

Bacteria are known as the main cause for disease outbreaks [1]. Defining an effective treatment requires characterizing the disease outbreak by identifying its pathogens. A bacterial type diagnosis is required, i.e. the identification of pathogens below the species level [1-2], for controlling the disease. Sub-grouping of bacterial species to bacterial types is used for many important pathogenic bacteria, such as the *Staphylococcus aureus* (*S. aureus*). The *S. aureus* species is a major cause of infections as well as farm animals' diseases such as mastitis of lactating cows [1], [3]. This bacteria tendency to develop resistance to antibiotics raises the importance of its identification via typing.

Phage typing is a method used for defining the types of a species via the species reactivity to a set of selected bacteriophages (phages) [1]. A phage is a bacterial virus activated by specific bacterial surface constituents of the checked species. The phage receptor binds to a bacterial surface component, invades and multiplies in the bacterial host. When a phage infects a layer of bacterial cells, a zone of lysis produces a plaque, viewed as a clear area in the bacterial lawn, such as the full circles (spots) in Figure 2(a). These represent positive reactions to different phages. When the phage receptor does not recognize any of the tested bacterial surface constituents, no plaque is formed and it is defined as a negative reaction. In this case no surface change is visible. The set of phages active against a culture of bacteria isolates, form a unique profile specific for each bacterial type. We term this profile the “phage profile”.

The identification of a bacterial type phage profile belongs to the general task of image array analysis. Analysis of image arrays is comprised of two important tasks: spot finding and spot analysis. Our research includes spot finding as well as spot categorization (labeling spots into positive and negative reactions) and phage profile extraction. A preliminary report of this work has appeared in [4]. To the best of our knowledge, no previous work has been done on automatic spot analysis for phage typing. Related work on spot finding has been described for microarrays and macroarrays, both on rigid slides and on flexible membranes. Many works on spot finding are found in the domain of cDNA microarray data analysis, where the goal is to identify the locations and extents of labeled DNA spots in a scanned microarray image [5], [6].

The spot finding task usually involves two objectives of image segmentation and grid positioning. A preliminary grid overlay [e.g. 5-8] separates the signal from the background. The grid partitions the image plane into windows, rectangular units uniformly spaced in an array overlaid on the image plane, such that each window contains a single reaction (spot). The windows are analyzed locally, each one separated into a spot region and a background region. Grid placement is commonly achieved with human intervention. Various methods have been used to segment each window, including histogram-based segmentation [8-10], seeded region-growing [11], shape-based segmentation [7, 12, 13] and more. Human intervention is important in most of the above-mentioned methods. Manual input consists of roughly circling the spot regions, a-priori setting the grid partitions and in determining parameter settings, such as intensity thresholds, shape and size of spots.

The proposed framework is comprised of the following main features: (1) A major focus of the work is on spot categorization and analysis, which was not previously done in the domain of phage typing; (2) The spot finding task is composed of global image segmentation into signal vs. background regions via unsupervised clustering, followed by a gridding procedure that provides localization of the individual spots. Note that most works in the field use localized processing only, and thus require a gridding process as a crucial first step of the system; (3) Statistical analysis of the spot region characteristics enables probabilistic categorization of the spot reactions and the transition from spot categorization to phage profiling per bacterial type.

In section 2 input data characteristics are presented. The methodology involving computer-vision and statistical modeling is presented in section 3. Experimental results on the *S. aureus* are shown in section 4. Discussion of the results is conducted in section 5.

2 Data Characteristics

Gray-level images of phage typing arrays are the visual input to the proposed system. The images are scanned using a UMAX scanner, Powerlook2 model, with a transparency adaptor. Each image is of size 532x532 pixels. An example of scanned images is presented in Figure 2(a). The petri-dishes seen in the images contain a surface of *S. aureus* species. Reactions to 60 different phages are present on the surface of the dish. The reactions are organized in a fixed array and known order. An *image-group* contains a set of images. A given database consists of image-groups, each group representing a particular *S. aureus* bacterial type.

A significant variability between the scanned images and irregularities in each image exist within a given database. Image contrast and dynamic range is considerably different across the image-group. Reaction shapes and sizes are irregular, both within an image as well as across the images. Reactions are not positioned in a uniform layout. Finally, the background, i.e. the dish surface, also exhibits non-uniformity due to inevitable differences in experimental conditions, and variability in the pigmentation of bacterial isolates.

3. Methods

Figure 1 presents the general framework proposed in this work. Visual array data is processed via two stages: a segmentation stage and a follow-up categorization stage. Statistical modeling via Gaussian Mixture Models (GMM) and Expectation-Maximization (EM) learning are utilized in both stages of analysis. A transition is made to phage profiling and the final output is a probabilistic signature of phage-reaction profile (phage profile) per image-group.

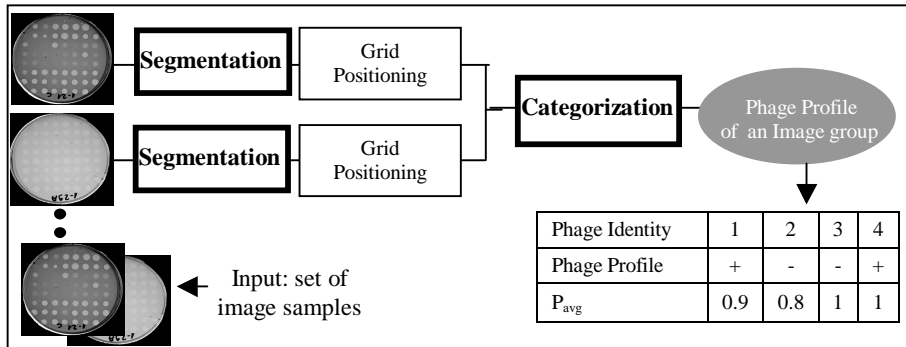


Fig. 1. Statistical analysis framework: from the visual array to the phage profile.

3.1 Statistical Modeling

In modeling image data, an initial transition is made from the raw pixel input to a selected feature space. Each pixel is represented by a feature vector and the image as a whole is represented by a collection of feature vectors. The underlying assumption is that a mixture of Gaussians generates the image features' distribution. The distribution of a random variable, $x \in \mathbb{R}^d$, is a mixture of k Gaussians if its density function is:

$$f(x|\theta) = \sum_{k=1}^K \alpha_k \frac{1}{\sqrt{(2\pi)^d |\Sigma_k|}} \exp\left\{-\frac{1}{2}(x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k)\right\}, \quad (1)$$

such that the parameter set $\theta = \{\alpha_k, \mu_k, \Sigma_k\}_{k=1}^K$ consists of:

$$\alpha_k > 0, \quad \sum_{k=1}^K \alpha_k = 1$$

$\mu_k \in \mathbb{R}^d$, Σ_k is a $d \times d$ positive definite matrix

where α_k is the prior probability for Gaussian k , and μ_k, Σ_k are the mean vector and covariance matrix of Gaussian k , respectively.

Given a set of feature vectors X_1, \dots, X_N the maximum likelihood estimation of θ is:

$$\hat{\theta}_{ML} = \underset{\theta}{\operatorname{argmax}} f(X_1, \dots, X_N | \theta) = \underset{\theta}{\operatorname{argmax}} \prod_{n=1}^N \sum_{k=1}^K \alpha_k f(X_n | \theta) \quad (2)$$

The EM algorithm is an iterative method to obtain the parameter set θ_{ML} increasing the likelihood function in each iteration [14].

The probabilistic affiliation of feature vector X_n to cluster (Gaussian) k is given by:

$$P(\text{label}(X_n) = k) = \frac{\alpha_k f(X_n | \mu_k, \Sigma_k)}{f(X_n | \theta)}. \quad (3)$$

Each feature vector, x_n , is assigned to the most probable Gaussian cluster, i.e. to the component of the model that maximizes the a-posteriori probability:

$$\text{Label}(X_n) = \underset{k}{\operatorname{argmax}} P(\text{label}(X_n) = k). \quad (4)$$

3.2 Image Segmentation: Signal vs Background

The objective of the segmentation phase is to extract a probabilistic separation of the data, per image, into the spot region and the background region. The segmentation is

performed globally on the entire image plane (rather than window by window). The segmentation task is treated as an unsupervised clustering task using the intensity feature, with two main clusters to be found. The image is represented by a mixture of two Gaussians ($k=2$), one Gaussian represents the image signal, the other represents the image background intensity distribution. The choice of k is based on apriori analysis of the data at hand. Once the model is learned, the probabilistic affiliation of each feature vector, the image pixel intensity, with each of the Gaussians in the model ($k=1,2$) can be computed (equation (3)), and each pixel of the original image, X_n , is then affiliated with the most probable Gaussian cluster (equation (4)). The result of the segmentation phase is a binary image in which an estimate of the signal region is given.

3.3 Grid Positioning: From the Image Plane to Image Spots

A transition is made from the global image signal to localized reactions, or spots. Extracting local areas of interest in an input image is accomplished by partitioning the data into windows via a grid. In this work we position the grid automatically based on the segmentation results. The grid-positioning algorithm is as follows: Starting with its original state, each binary (segmented) image is rotated in intervals of 1° between $+10$ and -10 degrees. For each angle of rotation, \mathcal{G} , a projection (sum) over the x-axis and the y-axis is computed. Thus, for each \mathcal{G} we get two projection functions, $f_{\mathcal{G}}(x)$ and $f_{\mathcal{G}}(y)$ for the x-axis and the y-axis, respectively. The angle α by which the image has to be de-rotated is determined by the angle for which a maximum projection value is found:

$$\alpha = \arg \max_{\theta} [f_{\theta}(x), f_{\theta}(y)]; \quad \mathcal{G} \in [-10\dots+10] \quad (5)$$

The set of maximal-strength signal values is extracted from the x and y-profiles of the de-rotated image. The location, for which the maximum projection value for the x-axis and the y-axis is achieved, serves as an anchor point from which the grid is defined. The grid interval is calculated as the average maximal-strength signal values. Utilizing the grid, spot finding is accomplished; a transition is made to local image analysis and further spot processing.

3.4 Spot Categorization

A categorization of each spot to positive or negative reaction is pursued. A set of feature vectors extracted per image-group is analyzed statistically for modeling positive and negative reactions. Using the learned model a categorization of each spot into the positive and negative clusters is enabled. A reaction is defined as positive, when there is a sufficient number of pixels of high-intensity present and a circular structure is evident. A transition is thus made from the raw pixels to a feature space that accommodates the signal strength and morphology criteria. The labeling task

becomes a clustering task within the selected feature space. The following features are extracted from each spot:

(1) Normalized Area (NA). The area of a spot, A , is the sum of pixels that comprise the spot. This sum is normalized by the average spot size within the image. For the average, only spots with an area above a certain threshold, T , are considered:

$$\frac{A - \text{Avg}(A > T)}{\text{Avg}(A > T)} * 100, \quad (6)$$

with the threshold value T empirically set to 200. The normalization factor is important in order to achieve invariance to the variability in the positive reaction area across the images in the dataset. The normalized area feature is therefore an estimate of the relative size of a reaction per spot.

(2) Shape Index (SI) = $\frac{4\pi A}{P^2}$, (7)

where A is the area of signal and P is the perimeter of the signal (sum of edge pixels per spot). This feature is used in the literature as a measure of circularity [15]. High values of SI represent spots with higher circularity and less graininess.

Feature vectors, X_1, \dots, X_N , ($X_n = (NA, SI)$) are extracted from all spots, $n=1 \dots N$, across all the images in a given image-group. Clustering of the features is pursued using GMM and EM. A two-cluster partition is used ($k=2$) to separate the space into the positive and negative reaction categories. The choice of a two class partitioning is motivated from the biological categorization of the spots into two reaction groups. Utilizing the learned GMM model, the probabilistic labeling of each spot is enabled (equation (3)). The affiliation of each spot to the most probable Gaussian cluster (equation (4)) produces image spot categorization.

3.5 From Spots to Phage Profiling

We next shift from the level of spot categorization to the level of phage profiling per image-group. Each phage (G_i) is probabilistically affiliated with the positive and negative reaction categories (k), by averaging across the corresponding spot probabilities:

$$P_{\text{avg}}(\text{label}(G_i) = k) = \frac{1}{N} \sum_{n \in G_i} P(\text{label}(X_n) = k), \quad (8)$$

with the averaging performed over the spots, X_n , $n=1 \dots N$, related to the same phage, G_i , for all images in the image-group. The phage label is determined by the higher average probability of the two:

$$\text{label}(G_i) = \arg \max_k P(\text{label}(G_i) = k). \quad (9)$$

A standard-deviation measure is computed over the spot probabilities to estimate the variability in the spot reactions within the image-group. The phage profile is taken as a collection of phage labels along with the average probabilities per phage, extracted in a pre-defined order across the image array.

4 Experiments and Results

We present experimental results for both spot and phage level analysis. The dataset used consists of 4 image-groups, each from a different farm. Each farm corresponds to a particular bacterial type. Three image-groups (#1, #2, #4) consist of 40 images. Group #3 contains 260 images. A central sub-array of 6*6 (36) phages is analyzed per petri-dish image, to avoid edge effects. Four petri-dishes are input at a time, thus a total of 144 (36*4) phages are considered in a phage profile.

The segmentation and gridding processes are exemplified in Figure 2. The segmented images in (b), are shown following de-rotation. Note that pixels that are affiliated with the “signal” Gaussian are displayed in white, while pixels that are affiliated with the “background” Gaussian are displayed in black. The separation between the signal and background regions is evident. The graininess present in the segmented images is due both to the low contrast input (as in image #1) as well as to biological reactions (as in image #2). The dark-particle noise evident in image #2 is removed in the segmentation process while the bright noise artifacts are interpreted as signal. Results of the gridding process and the transition to image spots are shown in Figure 2(c).

Spot analysis and categorization (section 3.3) is demonstrated in Figure 3 and Figure 4. The GMM learned from all feature-vector samples, using the shape index (SI) and normalized area (NA) features, for a particular image-group, is shown in Figure 3. A clear separation between the two major modes is evident, with one cluster (Gaussian) representing positive reactions and the second cluster representing the negative reactions. The spread of each cluster defines the variance in each group.

Spot *categorization* is achieved by computing the probabilistic spot affiliation to each of the Gaussians of the learned GMM, and then determining the most probable Gaussian cluster (equations (3) and 4)). Figure 4 presents an example of image with spot categorization. The reaction category is indicated (+/-), along with its probability. Spots that are perceptually very clearly categorized (into positive or negative reactions) are supported with higher probabilities than spots for which the reaction category is visually questionable.

Table 1. Correlation and statistical results of expert-based categorization versus automatic categorization

Category	True Positive	False Negative	False Positive	True Negative
Number of spots	5100	105	340	2807

Sensitivity (True Positive Categorization)	98%	Specificity (True Negative Categorization)	89%	Correlation with Supervised Categorization	95%
Positive Predictive Value	98%	Negative Predictive Value	96%		

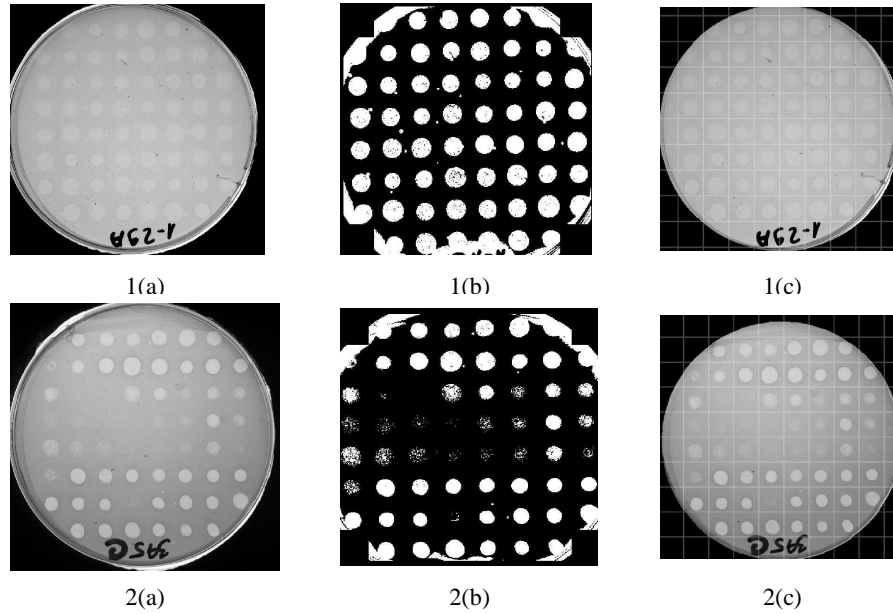


Fig. 2. Examples for segmentation and gridding: (a) Original images; (b) Segmented images; (c) Images overlaid with grid (de-rotated)

Prior to proceeding to the phage-level analysis, we wish to validate the spot categorization results. The validation process included a comparison of a large set of categorized spots to human-expert labeling as ground-truth. Results are given in Table 1 for 7992 spots, which were ascribed a label of positive and negative reactions by an expert. The algorithm achieved a correlation of 95% with the supervised categorization. A 98% correlation was achieved with the supervised positive categorization (sensitivity) and an 89% correlation was achieved with the supervised negative categorization (specificity). The positive predictive value, i.e. the probability that the spot reaction is manually categorized as positive when the automatic categorization is positive, is 98%. The negative predictive value is 96%.

4.1 Phage Profiling and Analysis

An example of a phage profile is shown in Table 2 (equations (8) and (9)). The categorization of each phage is given in the second row. The corresponding average probability, P_{avg} , and standard deviation (*std.*) are listed in the third and fourth rows, respectively. A high average probability indicates a similar (and strong) spot reaction in the particular image location, across the images in the image-group. For example, consider phage #7 in Table 2. The category is the negative category (-), P_{avg} is 100

with a std. of 0. We can conclude from this that all spots in the image-group have a negative response at 100% probability. A low average probability indicates either that the spot reactions (per phage) are not similar, which will be evident in a large *std.* value, or that for each labeled spot, the affiliation probability is low (equations (3) and (4)). Phage #14 in Table 2, for example, has a P_{avg} of 50% and a large *std.* The spot reactions for this phage are in fact split, with half of the spots having a large probability for the positive reaction and half having a large probability for the negative reaction.

Table 2. An example of a phage profile (19 out of 144 phages). Each phage is identified by a serial number and is affiliated with a reaction type (+ or -). The average probability for the reaction type is shown, along with the corresponding standard deviation

Phage Identity	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Phage Profile	+	+	+	+	+	+	-	-	+	+	-	+	-	-	-	-	-	+	-
Pavg	84	95	91	94	94	99	100	97	83	67	81	78	79	50	100	85	60	98	53
STD	0.37	0.21	0.28	0.23	0.24	0.05	0.00	0.18	0.38	0.47	0.39	0.42	0.41	0.50	0.00	0.36	0.49	0.12	0.50

The phage profiles are used in this work as a basis for analyzing similarities and differences of bacterial types. An assumption that serves as the ground-truth, is that an image-group corresponds with a particular bacterial type. In validating the phage profiles generated by the proposed system, profiles that are extracted from image-groups of a similar bacterial type are expected to have a similar signature, while phage profiles that are extracted from image-groups of different bacterial types, are expected to show large profile variations. The comparison between phage profiles is measured as the percentage of similar categorization, (PSC) between the phages.

Two experiments are conducted to validate the extracted phage profiles. In the first experiment, a given image-group of 260 images is divided into seven image subgroups. A phage profile is generated for each image-subgroup and PSC values are computed across the profiles. Utilizing the full profiles (144 phages) results in a PSC value across the seven subgroups of 78%. Close examination reveals that phages for which the reaction category is not consistent across the profiles, have low average probabilities. Thresholding the phage profiles, using an average probability threshold value of 80%, results in a reduced size profile of 96 phages. Within this phage-profile set, the PSC is 100%, i.e. a full similarity is achieved across the seven profiles.

In a second experiment, phage profiles for the four image-groups in the dataset are investigated. Table 3 presents the phage profiles generated for the four image-groups. The phage profiles indicate a difference between groups #1, #3 and #4, with a similarity between groups #2 and #3. Figure 5 displays corresponding PSC values. The similarity percentage is plotted vs the percentage of profile thresholding. Three curves are shown, each representing PSC values for a particular pair of image-groups. The curve representing groups #2 and #3 has a high percentage of similarity of above 90% without any thresholding, and an increased value of 100% following profile thresholding. A similarity percentage of less than 60%, without thresholding, is seen

in the comparison between groups #1 and #3, and groups #4 and #3. The distinct groups remain mostly around the 60% and 70% range throughout. From these results, it may be concluded that different bacterial types are present in groups #1, #3 and #4. The same type is most likely present in groups #2 and #3.

5 Conclusions

In this work we developed an automatic tool which analyses microbiological data and extracts bacterial types as probabilistic phage profiles. Image array analysis is used to analyze spots by advanced computer vision algorithms, with minimal human intervention. Automation of biological analyses is of importance for efficient and accurate research and production, especially as the amount of data is constantly increasing.

In the current work visual array data was segmented and categorized utilizing statistical imaging methods (GMM and EM). The output of the system is a probabilistic phage profile representing the input image-group. Supervised validation was used in the spot categorization task. Strong correlation, of 95%, was found with the human expert labeling (Table 1).

An important objective of the work is the ability to identify similarity and distinction amongst phage profiles within and across image-groups. Examples of phage profiles are shown in Tables 2 and 3. In the first experiment conducted, PSC values reached 100% for a comparison across subgroups, i.e. originating from a single image-group. Figure 5 shows the percentage of similar phage categorization, for different pairs of image-groups. The PSC is seen to correspond with the given ground-truth. These results are encouraging in that the automated extracted phage profiles seem to be able to validate hypothesis about bacterial types present in a given dataset.

Our study suggests a generic tool that aids the microbiologist in transforming and supplementing data into useful information for analysis. An objective and consistent processing is provided. The expert can effect the evaluation by determining a (optional) probability threshold, for the compaction of the representative profile into the set of the more probable profile reactions. Probabilistic phage profiling can provide a strong basis for further analysis and bacterial type classification. The methodology presented includes general processing steps that may be applicable regardless of scale: from the micro-array to the macro-array. It should be noted, that in each case, adaptation is needed to the data characteristics, and the required level of accuracy. Related domains include phage therapy, a developing domain related to drug discovery, and the domain of cDNA microarray data analysis.

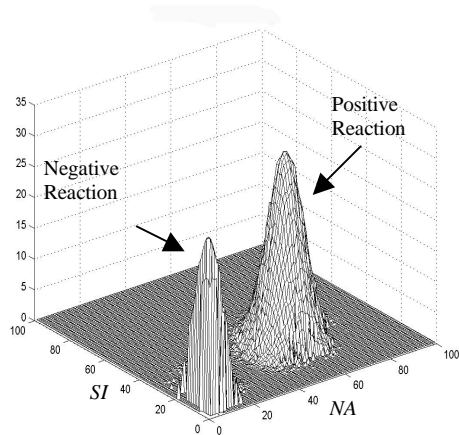


Fig. 3. The learned GMM for the input data. The two modes represent positive and negative reactions, respectively

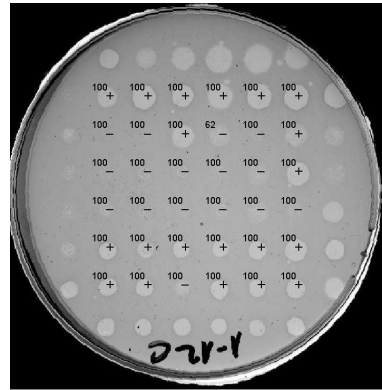


Fig. 4. Spot categorization along with corresponding probabilities

Table 3. Profiles (19 out of 144 phages) from four image-groups

Phage no.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Group #1	-	-	-	+	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-
Group #2	+	+	+	+	+	+	-	-	+	-	-	-	-	-	-	-	-	+	-
Group #3	+	+	+	+	+	+	-	-	+	+	-	+	-	-	-	-	-	+	-
Group #4	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+

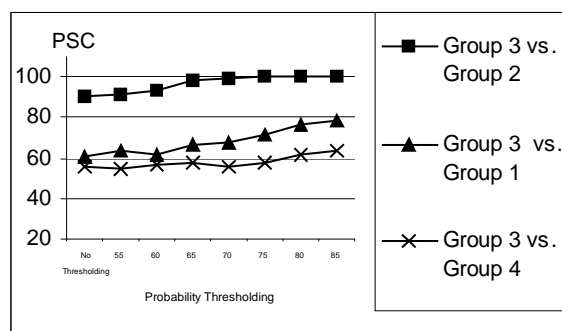


Fig. 5. The PSC (percentage of similar categorization) in different pairs of image-groups vs the percentage of profile thresholding

References

1. Emori T.G, Gaynes R.P.: An Overview of Nosocomial Infections, including the Role of the Microbiology Laboratory. *Clinical Microbiology Reviews* 6 (1993) 428-442
2. Tenover F.C., Arbeit R.D., Goering R.V.: How to Select and Interpret Molecular Strain Typing methods for Epidemiological Studies of Bacterial Infections: a Review for Healthcare Epidemiologists. *Infection Control and Hospital Epidemiology* 18 (1997) 426-439
3. Spring Diagnostics. <http://www.itek.co.il/spring>
4. Trattner S., Greenspan H., Teper G, Abboud S. Automatic Identification of Bacterial Types Using Statistical Imaging Methods. *Proceedings of SPIE International Symposium on Medical Imaging, San Diego, USA (2003)*.
5. Yang Y.H., Buckley M.J., Dudoit S., Speed T.P.: Comparison of Methods for Image Analysis on cDNA Microarray Data. *Journal of Computational and Graphical Statistics* 11(1) (2002) 108-136
6. Smyth G.K., Yang Y.H.: Statistical Issues in cDNA Microarray Data Analysis. In *Functional Genomics: Methods and Protocols*. Brownstein M.J. and Khodursky A.B., Eds., *Methods in Molecular Biology series*, Humana Press, Totowa, NJ (2003)
7. Ideker J., Haynor T., D.: Dapple: Improved Techniques for Finding Spots on DNA Microarrays. *University of Washington CSE Technical Report UWTR (2000)*
8. Chen Y., Dougherty E.R., Bittner M.L.: Ratio Based Decisions and the Quantitative Analysis of cDNA Microarray Images. *Journal of Biomedical Optics* 2 (1997) 364-374
9. QuantArray Analysis Software. <http://lifesciences.perkinelmer.com>
10. Scanalytics MicroArray Suite. <http://www.scanalytics.com>
11. Adams R., Bischof L.: Seeded Region Growing. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 16 (1999) 641-647
12. Eisen M.B., ScanAlyze User Manual. *Stanford University, Palo Alt*, <http://rana.lbl.gov> (1999)
13. Wang X., Ghosh S., Guo S.W.: Quantitative Quality Control in Microarray image Processing and Data Acquisition. *Nucleic Acids Research* 29(15) (2001) e75
14. Bishop C.M.: *Neural Network for Pattern Recognition*. Clarendon Press, Oxford (1996)
15. Sonka M., Fitzpatrick J.M.: *Handbook of medical imaging, Vol. 2: Medical Image Processing and Analysis*, SPIE Press, Washington (2000)