MULTIVARIATE ANALYSIS OF VARIANCE APPLIED TO MICROBIOLOGICAL EXPERIMENTAL DATA

Michele Wolf

May 10, 2003
Master's Paper

MULTIVARIATE ANALYSIS OF VARIANCE APPLIED TO MICROBIOLOGICAL EXPERIMENTAL DATA

ABSTRACT

Multivariate analysis of variance (MANOVA) was used to analyze a microbial data set containing variables quantifying biofilm structure. The data were obtained from a laboratory experiment designed to evaluate differences between two strains of bacteria in two different environments. A linear combination of response variables, an index, was developed that significantly distinguished among the four combinations of strain and flow.

INTRODUCTION

We live in a complex world, where few individuals, if any, live in isolation. For the most part, we are surrounded by communities that exhibit a variety of interconnections among living things. These connections are often numerous and complicated. Statistical analyses of data collected on such systems, allows us to identify patterns in nature and disentangle the delicate interrelationships found within these systems.

In the "real world," data collected from complex systems are inherently multivariate in nature. Analyzing such data is a daunting task because much of the data contain multiple measurements on the same individuals, thus measurements are not independent of one another. Moreover, inter-correlations among measurements are almost certain. Unfortunately, inter-correlations are often not addressed by researchers,

most prefer to use the univariate statistical framework for data analysis when the multivariate framework is often more appropriate for this kind of data. Univariate approaches applied to multivariate data are common practice. However, conclusions based on univariate analyses may be in error when differences among treatments are due to chance alone. For example, when separate univariate analyses of variance are performed on multiple response variables, these multiple univariate tests suffer from increased family wise error rates. Thus the best approach, is to analyze multivariate data using multivariate techniques.

Microbial biofilms are micro-organisms that are surrounded by polysaccharide secretions, hence the name biofilms. Biofilms can be found coating a multitude of surfaces from rocks in streambeds to catheter tubes, they readily adhere to any surface covered by water (Costerton 2001). Their presence is at best innocuous and at worst can cause infections or cause millions of dollars of damage to commercial products like stainless steel piping. Biofilms require very little nutrients and oxygen, so they are able to live in a variety of habitats, from slow moving to high velocity channels. Water not only supplies required nutrients, oxygen and secondary colonizers, but also removes waste and by-products from the biofilm.

Biofilm researchers use images captured by a digital camera on a high-powered microscope to acquire basic information about biofilms. Data extracted from biofilm images are lengthy, complex, arduously obtained and inherently multivariate. Such data often contain multiple variables collected on a single biofilm image. In the investigation discussed below, measurements were performed on each image to quantify the morphological structure of the biofilm. Although much work has been performed on

free-floating bacteria, few studies have examined sedentary bacteria until now. Moreover, an attempt at quantifying morphological structure of sedentary bacteria via image analysis is in its infancy. Data arising from such research fall into the realm of the multivariate framework for statistical analysis.

The process of distinguishing biofilms based on morphological characteristics can be difficult. Until recently, choosing characteristics that adequately describe the shape, size and form of the biofilm has been mainly qualitative. Attempts to quantify biofilm structure have mostly been limited to fractal dimension (Yang et *al.* 2001). In an effort to go beyond the limitations of qualitative descriptions and fractal dimension descriptions, quantitative measures were obtained from image analysis of biofilms developed through designed experiments (Lewandowski et *al.* 1999). These measures were derived to capture the underlying processes as well as the structure of the biofilm.

Stoodley et.al. (2001; 1999b) and Yang et. al. (2001) have developed a more comprehensive set of variables to quantify biofilm structure via image analysis. The overall objective of this paper was to use a multivariate approach (MANOVA) to evaluate microbial biofilm structure. The specific goals pertaining to the analysis of this data set were three-fold. Firstly, because the data contain multiple variables describing each image, we will use multivariate analysis of variance, MANOVA, to analyze the data. Secondly, we will determine which of the variables were most important in distinguishing among bacterial strains in different environments and retain those variables for further analysis. Lastly, we will create an index that can potentially be used to measure morphological differences among bacterial strains.

MATERIALS AND METHODS

A designed experiment was conducted using two strains of *Pseudomonus areuganosa* in two flow regimes (Stoodley et. al., 1999). Experimental flow chambers were prepared such that each contained slides in the bottom of a tank. The water was inoculated with the particular biofilm strain of interest and allowed to develop for a period of time. At set time intervals, *in situ* digital images were acquired. Images were analyzed, and data for 12 variables recorded. A detailed description of experimental procedures may be found in Stoodley et. al. (1999).

The data set resulting from experiments performed on two strains of *Pseudomonus areuganosa* contained information on 12 response variables and two explanatory variables. Porosity, horizontal run length and vertical run length were three response variables designed to capture the overall size of the biofilm. Porosity (Lewandowski 2002) is the ratio of void area to total area, so it is the amount of the surface that is not covered by biofilm, so small values indicate a large portion of the image is covered by a biofilm. Horizontal run length measures the expected dimension of a cell cluster in the horizontal direction while vertical run length measures the expected dimension of a cell cluster in the vertical direction. Therefore both are measures of cluster size. Other response variables, such as fractal dimension, textural entropy, and homogeneity, captured the roughness of the biofilm boundary and the randomness of the grayscale image. Finally, response variables such as average diffusion distance, maximum diffusion distance, aspect ratio, contrast, correlation, energy and autoregressive structure characterize both the size of the clusters and their general shape.

The factors for this experiment were strain and flow velocity, each having two levels (i.e. two strains and two flow rates).

The multivariate technique MANOVA is a generalization of analysis of variance (ANOVA) that allows the researcher to analyze more than one response variable at a time (Barker 1984; Bray 1985; Collins 1980). As an extension of ANOVA, MANOVA is designed to distinguish among group means using several response variables rather than a single response variable. Furthermore, this technique allows the researcher to look at the relationship among all the variables *simultaneously* rather than looking at each in isolation, while controlling for the intercorrelations among the response variables. The advantage to using MANOVA is that the analyst looks at the multi-dimensional variable space, trying to elucidate group differences in p-space that may not be apparent in each space individually. Thus, MANOVA may be a more powerful test than doing separate one-way ANOVA's.

The assumptions (Bray 1985) needed for MANOVA parallel those needed for ANOVA.

1. The units are randomly sampled from the population of interest
2. The observations are independent of one another
3. The response variables have a multivariate normal distribution within each group
4. The p groups have a common within-group population covariance matrix

The experimenters followed a spatial sampling protocol that provides credibility to assumptions one and two. Assumptions three and four were evaluated during the data exploration phase of the analysis. There were not enough observations to evaluate the

multivariate normality assumption; therefore, we only checked the marginal normality of each variable separately.

RESULTS

Univariate analyses (Table 1) were performed on each of the response variables to determine if the factors or interaction terms were significant and to obtain residuals for purposes of evaluating assumptions. Normality was assessed using the Anderson Darling test, while non-constant variance was assessed using residual plots (Table 1). Three of the variables, Porosity, Homogeneity, and Contrast conformed to the normality and constant variance assumptions. However, Energy, TE and FD had non-constant variance and ADD, MDD, HRL, VRL, Correlation and AR were both non-normal and had non-constant variance. Transformations were performed in an effort to remedy violations of the MANOVA assumptions. The variables ADD, MDD, HRL, VRL, AR and Energy were log (base 10) transformed while TE was exponentiated to try and make them more normal and less heteroscedastic. Normality tests and residual plots were again performed on each of the transformed variables to assess violations of MANOVA assumptions. Results indicate some variables still violated MANOVA assumptions (Table 2). Correlation matrices were constructed to determine redundancies within the set of response variables. Variables containing similar information were culled from the data set.

Response variables were reduced from 12 to 6 based on extensive exploratory analysis. The sets of variables, HRL and VRL, as well as MDD and ADD, contain very similar information, thus one from each set (VRL and ADD) was eliminated. Further

analysis revealed Homogeneity was proportional to contrast, so it was redundant and therefore eliminated from the analysis. Energy and FD were highly correlated with TE (-0.954 and 0.804, respectively) so they were also eliminated from the analysis. Surprisingly, initial MANOVA analyses showed that Contrast was exactly a linear combination of other responses; therefore, it was also removed from the analysis. The 6 remaining response variables were Porosity, logMDD, logHRL, logAR, log(Corr-4) and expTE. These variables were standardized and the MANOVA performed on the standardized variables. There were two outliers in this data set. While both were real observations, not data entry errors, they were eliminated from the data set for the following reasons. When looking at Porosity values, we could see that the value for observation 93 was high (0.995101). This indicated that there was a high ratio of void area to total area. In essence, there was no biofilm on the slide and so all of the measurements were made on an empty slide. The values for this observation were therefore extremely different from all of the other observations. This point was removed from the analysis. Similarly, LogCorr had a very large outlier in the negative direction. All of the correlation values were close to four, except observation 91 which had a very small value (0.00047). This value indicated excessive noise in the image, so it was eliminated from the analysis.

Multivariate analysis of variance was performed on the data using the reduced data set. Response variables for the model were Porosity, logMDD, logHRL, logAR, log(Corr-4) and expTE, while explanatory variables included strain, flow and the interaction between strain and flow. Results from the overall MANOVA test showed highly statistically significant flow and strain effects (Table 3; p-value < 0.003) and

almost statistically significant interaction effects (Table 3; p-value = 0.06). A single factor was then created coding for each of the four treatments (strain 1 flow 1 = treatment 1, strain 1 flow 2 = treatment 2, strain 2 flow 1 = treatment 3, strain 2 flow 2 = treatment 4). A MANOVA was again performed on the data using the six response variables and the single treatment factor. The treatment effect was highly statistically significant (Table 4; p-value < 0.0001).

The largest eigenvalue (1.9075, Table 5) and associated eigenvector accounted for much of the separation among the treatments (Table 5, cumulative proportion 0.8925). The first eigenvector was used to create Index1, a linear combination of the response variables that discriminated among treatments (equation 1).

$$\text{Index1} = 0.02606(\text{Sporo})+0.12627(\text{SLogHRL})+0.06915(\text{SlogAR}) -0.00247(\text{SLogMDD}) -0.02952(\text{Slog(Corr-4)})-0.02372(\text{SexpTE}) \quad (1)$$

An ANOVA was performed with Index1 as the response and strain, flow and the interaction between the two as the explanatory variables. Results were statistically significant for strain (Table 6, p-value = 0.012) and flow (Table 6, p-value < 0.0001), but were not statistically significant for the interaction (Table 6, p-value = 0.558, Figure 1). These results indicate that the interaction is not needed when trying to distinguish among strains and flows using Index1, therefore it was not included in further analyses.

Index1 was simplified by converting the coefficients to whole numbers and dropping SlogMDD because of its small coefficient. This simplification, called General

index1 makes the relationships among the variables more apparent and easier to interpret (equation 2).

General index1 = 3(Sporo)+13(SlogHRL)+7(SlogAR)-3(Slog(Corr-4))-3(SexpTE). (2)

The General index1 was submitted to ANOVA to ensure it retained the ability to distinguish among strains and flow regimes. Results for the general index were similar to those for Index1 (Table 7). Both strain and flow were statistically significant (p-values of 0.0132 and < 0.0001, respectively). Again, the interaction between strain and flow was not statistically significant (p-value = 0.543, Figure 2). The plot of the general index (Figure 3) clearly illustrates the separation between the two flow regimes. Points in red are from the turbulent flow, whereas points in black are from the laminar flow. The separation between strains is not as apparent, however, the squares represent the wild type strain, *Pseudomonus areuganosa*, while the circles represent the mutant type derived from *Pseudomonus areuganosa*. Visually, there are more squares below the x-axis than above.

Using the general index, four points were chosen (Figure 3) to evaluate the morphologies of the biofilms the index claimed were different. Biofilm images used for data analysis are presented in figure 4. Morphologies among these four images are distinct.

DISCUSSION

Using MANOVA techniques, an index was developed to distinguish turbulent flows from laminar flows and wild type morphologies from mutant type morphologies. It

was not obvious from univariate analyses that the biofilm morphologies for the two strains were significantly different. This statistical significance was an important result to the microbiologists. The index captured the size, shape and structural complexity of the biofilm grown under controlled experimental conditions. The index is mathematically straightforward lends itself to ease of use and interpretation. It is general enough that for similar experimental scenarios it may be useful to distinguish among bacterial strains and flow regimes. However, its use should be limited to similar experimental scenarios. Future work with this index would include using it on a new data set that contains the same variables used in this analysis, and evaluating its ability to separate groups in the new data set. It is of great interest to determine whether the coefficients for the index are repeatable.

LITERATURE CITED

Barker, Harry R. and Barker, Barbara M. Multivariate Analysis of Variance (MANOVA) A Practical Guide to Its Use in Scientific Decision Making. 1984. The University of Alabama Press, Alabama, 127pp.

Bray, James H. and Maxwell, Scott E. 1985. Multivariate Analysis of Variance. Sage Publications, London, 80pp.

Collins, C. Chatfield. 1980. Introduction to Multivariate Analysis. Chapman and Hall, London, 246pp.

Costerton, J. William and Stewart, Philip S. 2001. Battling biofilms. Sci. Am. 285(1), 61-67.

Lewandowski, Z. Notes on biofilm porosity. 2002. Wat. Res. 34, 2620-2624.

Lewandowski, Z., G. Harkin, and H. Beyenal. 2002. Author's response. Wat. Res. 36, 807.

Lewandowski, Z., Webb, D., Hamilton, M. and Harkin, G. 1999. Quantifying biofilm structure. Wat. Sci. Tech. 39, 71-76.

Stewart, Philip S. and Costerton, J. William. 2001. Antibiotic resistance of bacteria in biofilms. Lancet 358, 135-138.

Stoodley, P., A. Jacobsen, B. C. Dunsmore, B. Purevdorj, S. Wilson, H. M. Lappin-Scott, and J. W. Costerton. 2001. The influence of fluid sheer and $AlCl_3$ on the material properties of *Pseudomonas aeruginosa* PAO1 and *Desulfovibrio* sp. EX265 biofilms. Wat. Sci. Tech. 43, 113-120.

Stoodley, P., P. F. Jorgensen, P. Williams, and H. M. Lappin-Scott. 1999a. The role of hydrodynamics and ahl signaling molecules as determinants of structure of *Pseudomonas aeruginosa* biofilms, p. 323-330. *In* R. Bayston, *et al.* (ed.), Biofilms: The Good, the Bad, and the Ugly. J.W.T. BioLine, Cardiff, UK.

Stoodley, P., Z. Lewandowski, J. D. Boyle and H. M. Lappin-Scott. 1999b. The formation of migratory ripples in a mixed species bacterial biofilm growing in turbulent flow. Env. Microbiol. 1, 447-455.

Yang, X., H. Beyenal, G. Harkin and Z. Lewandowski. 2001. Evaluation of biofilm image thresholding methods. Wat. Res. 35, 1149-1158.

Yang, X., H. Beyenal, G. Harkin and Z. Lewandowski. 2000. Quantifying biofilm structure using image analysis. J. Micro. Meth. 39, 109-119.

Table 1 – Results from univariate analyses and evaluation of MANOVA assumptions.

Variable	Strain	Flow	Interaction	Normality	Variance
SPoro	NS	*	NS	NS	Constant
SHomo	NS	***	NS	NS	Constant
SCont	NS	***	NS	NS	Constant
SEnergy	NS	NS	NS	***	Constant
STE	NS	**	NS	***	Constant
SFD	NS	***	NS	*	Constant
SADD	NS	***	NS	***	Non-const
SMDD	NS	***	NS	***	Non-const
SHRL	NS	***	NS	***	Non-const
SVRL	NS	***	NS	***	Non-const
Scorr	NS	NS	NS	***	Non-const
SAR	***	***	NS	**	Non-const

Table 2 – Results from univariate analyses and evaluation of MANOVA assumptions, during the second analysis.

Variable	Strain	Flow	Interaction	Normality	Variance
SPoro	NS	*	NS	NS	Constant
SHomo	NS	***	NS	NS	Constant
SCont	NS	***	NS	NS	Constant
LogEnergy	NS	NS	NS	***	Constant
expTE	NS	***	NS	**	Constant
LogFD	NS	***	NS	**	Constant
LogADD	NS	***	NS	***	Constant
LogMDD	NS	***	NS	NS	Constant
LogHRL	NS	***	NS	NS	Constant
LogVRL	NS	***	NS	NS	Constant
Logcorr	NS	NS	NS	***	Non-constant
LogAR	***	***	NS	*	Non-constant

Table 3 – MANOVA results for initial evaluation of response variables.

MANOVA for flow s = 1 m = 2.0 n = 47.5

Criterion	TestStatistic	F	DF	P
Wilk's	0.36707	27.875	(6, 97)	0.000
Lawley-Hotelling	1.72424	27.875	(6, 97)	0.000
Pillai's	0.63293	27.875	(6, 97)	0.000
Roy's	1.72424			

MANOVA for strain s = 1 m = 2.0 n = 47.5

Criterion	Test Statistic	F	DF	P
Wilk's	0.81959	3.559	(6, 97)	0.003
Lawley-Hotelling	0.22013	3.559	(6, 97)	0.003
Pillai's	0.18041	3.559	(6, 97)	0.003
Roy's	0.22013			

MANOVA for flow*strain s = 1 m = 2.0 n = 47.5

Criterion	Test Statistic	F	DF	P
Wilk's	0.88557	2.089	(6, 97)	0.061
Lawley-Hotelling	0.12921	2.089	(6, 97)	0.061
Pillai's	0.11443	2.089	(6, 97)	0.061
Roy's	0.12921			

Table 4 – MANOVA results using reduced response variables and treatment as the predictor variable.

MANOVA for treatment s = 3 m = 1.0 n = 47.5

Criterion	Test Statistic	F	DF	P
Wilk's	0.27897	8.710	(18, 274)	0.000
Lawley-Hotelling	2.13721	11.359	(18, 287)	0.000
Pillai's	0.84750	6.496	(18, 297)	0.000
Roy's	1.90747			

Table 5 – Results from eigen analysis for treatment.

Eigenvalue	1.9075	0.2151	0.01460	0.00000	0.00000	0.00000
Proportion	0.8925	0.1007	0.00683	0.00000	0.00000	0.00000
Cumulative	0.8925	0.9932	1.00000	1.00000	1.00000	1.00000

Eigenvector	1	2	3	4	5	6
SPoro	0.02606	-0.0170	0.1204	0.1052	0.0669	-0.1690
SLogMDD	-0.00247	-0.1441	0.1102	-0.0700	-0.0040	-0.0067
SLogHRL	0.12627	0.1955	0.0619	0.1306	0.0245	-0.1534
SLogAR	0.06915	-0.0933	-0.0703	-0.0281	-0.0077	-0.0211
SLog(Cor	-0.02952	0.0451	0.0448	-0.0383	0.0641	-0.2176
SexpTE	-0.02372	0.0473	0.0864	0.0489	-0.0059	-0.3073

Table 6 – General linear model results using index 1 as the response and strain, flow and the interaction as predictor variables.

General Linear Model: Index1 versus strain_1, flow

```
Factor      Type Levels Values
strain_1    fixed      2 1 2
flow        fixed      2 1 2

Analysis of Variance for Index1, using Adjusted SS for Tests
Source          DF     Seq SS     Adj SS     Adj MS       F      P
strain_1         1    0.18867    0.06351    0.06351    6.48  0.012
flow             1    1.71533    1.70875    1.70875  174.30  0.000
strain_1*flow    1    0.00338    0.00338    0.00338    0.34  0.558
Error          102    0.99996    0.99996    0.00980
Total          105    2.90734
```

Table 7 – General linear model results using general index 1 as the response and strain, flow and the interaction as predictor variables.

General Linear Model: General Index1 versus strain_1, flow

```
Factor      Type Levels Values
strain_1    fixed      2 1 2
flow        fixed      2 1 2

Analysis of Variance for General, using Adjusted SS for Tests
Source          DF     Seq SS     Adj SS     Adj MS       F      P
strain_1         1     2035.5      687.0      687.0    6.44  0.013
flow             1    18596.5    18523.0    18523.0  173.51  0.000
strain_1*flow    1       39.8       39.8       39.8    0.37  0.543
Error          102    10888.8    10888.8      106.8
Total          105    31560.6
```

Figure 1 – Interaction plot for index 1.

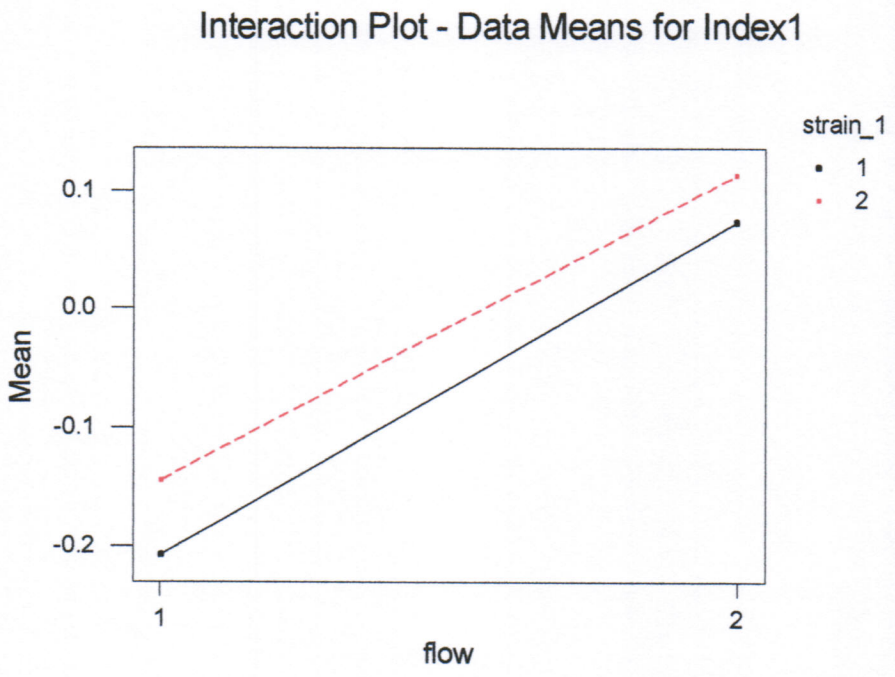

Figure 2 – Interaction plot for general index 1.

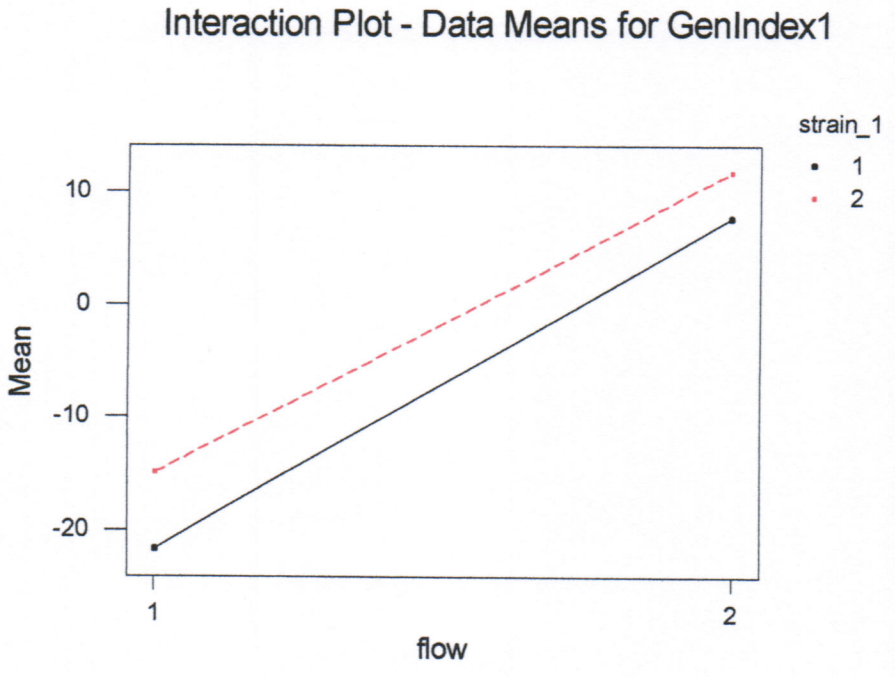

Figure 3 – Plot of data observations using general index 1.

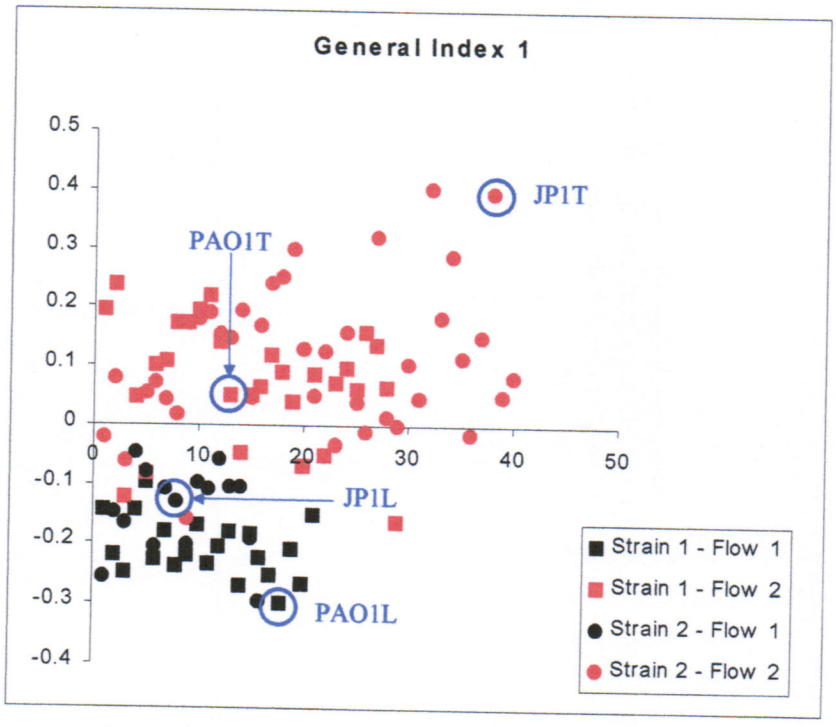

Figure 4– Digital images of biofilms obtained from general index 1.