Journal of Engineering and Applied Sciences
MEDWELL PUBLICATIONS

Statistical Analysis of Morphological Growth phases of Cyanobacteria

[1]Sabeeha Sultana and [2]Mohammad Basha
[1]*Department of Computer Science, PES University, Bangalore, Karnataka, India*
[2]*Kumar Oranics Pvt. Ltd., Jigani, Bangalore, Karnataka, India*

Key words: Cyanobacteria, Linear regression, Scatter plots, Correlation, Time series, Chlorophyll, Growth Phases, Statistics, Errorbars, Biomass, Training model

Corresponding Author:
Sabeeha Sultana
Department of Computer Science, PES University, Bangalore, Karnataka, India

Page No.: 6-17
Volume: 16, Issue 1, 2021
ISSN: 1816-949x
Journal of Engineering and Applied Sciences

Abstract: An automatic generic tool is developed to identify the morphological growth phases of microbiological data types using computer-vision and statistical modelling techniques. In algae phage (phage) typing, representative profiles of morphological growth stages of different algae types are extracted. Present systems rely on the subjective reading of the growth profiles by a human expert which is time consuming and prone to errors. The statistical methodology existing in this work, provides for an automated, objective and robust analysis of the visual image data, along with the facility to cope with increasing data volumes. Validation is performed by comparison to an expert manual segmentation and labelling of the growth phage profiles. The statistical analysis performed on time series data extracted is important for understanding relationships between parameters, provides insight to the growth curve of micro algae and cyanobacteria (correlation) and an essential step to forecast yield of biomass, etc. or predict the duration to achieve a certain yield of a pigment or protein, etc., for commercial applications. There are a number of methods for modelling time series data and being able to predict specific values; specifically, regression analysis and Analysis of Variance (ANOVA) are foremost among them. Computation of the correlation coefficient aids in better understanding the relationships that exist between various parameters that evolve with time and change with different phases of the growth of the organism (and cyanobacteria). This study focuses on statistical techniques for analysis of time series data.

INTRODUCTION

The branch cytology which deals with the study of cells in terms of their origin, structure, organelles and functional properties, is of key importance in biology and medicine. It consists of the recognition of cell types, fundamental for understanding biological differentiation. In this study, we investigate the feasibility of leveraging machine learning for morphological features to enable the identification of nine different algae types in an automated fashion. More specifically, we explore and investigate the efficacy of a number of different morphological and spectral fluorescence features extracted from multi-band fluorescence imaging data when used to train neural

network classification models designed for the purpose of identification of algae types in an automated manner. Manual identification of cyanobacteria according to correct taxonomy is generally impeded by obstacles such as declining the number of taxonomists and the increase in the number of described species, also discussed in detail by Gaston and O'Neill[1] and Soberon and Peterson[2] which makes identification of samplings a difficult and time consuming activity. Automated tools may significantly contribution in species recognition by facilitating reliable recognition of any specimens in a population. Automated methods that rely on pattern recognition and image analysis have been widely applied for recognition and categorization of biological images in the field of biodiversity[3-12]. Content based retrieval is one of the common text-based approaches in image retrieval domain[3-15] in a way that images of specimens are matched with images in data- base according to visual content (colour, shape, texture) similarities. For recognition of species, visual features that are extracted from digital images based on morphology and taxonomic information play a vital role.

A lot of work is carried by many by researchers in the literature on this subject. The statistical image analysis for automatic identification of bacterial types are proposed[16]. The artificial neural network approach for bacterial classification has been investigated[17]. The data mining techniques are employed for the classification of HEp-2 cells in Perner[18] in which a simple set of shape features are used for classification of bacterial cells. Wahlby *et al*.[19] have investigated algorithms for cytoplasm segmentation of fluorescence labelled cells using statistical analysis techniques based on shape descriptive features. A computer-aided system for the image analysis of bacterial in microbial communities using geometric shape features have been investigated in Liu *et al*.[20]. The automatic identification and classification of bacilli bacterial cell growth phases has been proposed in and (2010) using geometric shape features. A new image analysis tool to study biomass and of three major groups in an alpine lake using geometric features have been proposed in Posch *et al*.[21]. An efficient automated method for image-based classification of microbial cells has been investigated. Quantification of uncultured microorganisms by fluorescence microscopy and digital image analysis has been carried out[22]. The cell image analysis ontology using geometric and statistical features has investigated by Hiremath and Bannigidad[23].

MATERIALS AND METHODS

The slides and cover slips were prepared as: The slides and cover slips were thoroughly cleaned, dried and ensured of being free from dust, debris and grime because it touches the object being observed and has greater potential to contaminate the specimen if careful handling is not undertaken. The flat slide was placed on a clean, dry surface. A few drops of the sample were obtained using plastic pipettes (sample taken from a clear surface). A small amount is collected from the green area (sample taken from the bottom) with a pair of tweezers and placed on the centre of the slide. One drop of liquid sample was squeezed out onto the direct centre of the flat slide.

The cover slip was gently lowered onto the flat slide. One edge the cover slip was placed down first before lowering the rest. The cover slip must not be pressed down once it is in place. The slide and cover slip combination was picked up and gently placed on the viewing tray of the microscope.

RESULTS AND DISCUSSION

Let the horizontal or 'x-axis' represent the number of days and the vertical or 'y-axes represent the biomass variation on each of the days. We have a total of 15 readings recorded on alternate days through an entire month. The relationship between the number of days and amount of biomass produced is calculated. Person correlation is applied to the data set.

Analysis; Time series plot: The raw data tabulated contains two variables: 'x' (time stamp) and 'y' (a measurement of the parameter of interest). In all, there are 15 observations recorded through wet lab studies. A plot of the parameter against time helps us study the trend in the data.

Scatter plot: We plot the points on a graph to get a scatter diagram (i.e., the values of two parameters for corresponding time stamps). The scatter diagram helps us understand the relationship between the two parameters: whether the data is uncorrelated or the correlation is positive or negative.

Time series plot: In our study, there is a positive correlation (observed as an upward trend in the plot) till a certain point of time (as the number of days increases the value of either biomass/chlorophyll) and then we see downward trend, a negative correlation is observed after a certain reading. This indicates that the value of variable 'x' that is a number of days increases the value of the biomass/chlorophyll a decrease as shown in Fig. 1.

Figure 2 shows the production of chlorophyll a and b for the following species: *Pediastrum* sp., *Chlorella* sp., *Scenedesmus* sp., *Scenedesmus quadriquada* sp. and *Chlorococcum* sp. We note that the yield of chlorophyll b is more than compared with the production of chlorophyll a. Further, both chlorophyll b,a increases exponentially during the exponential phase and slowly

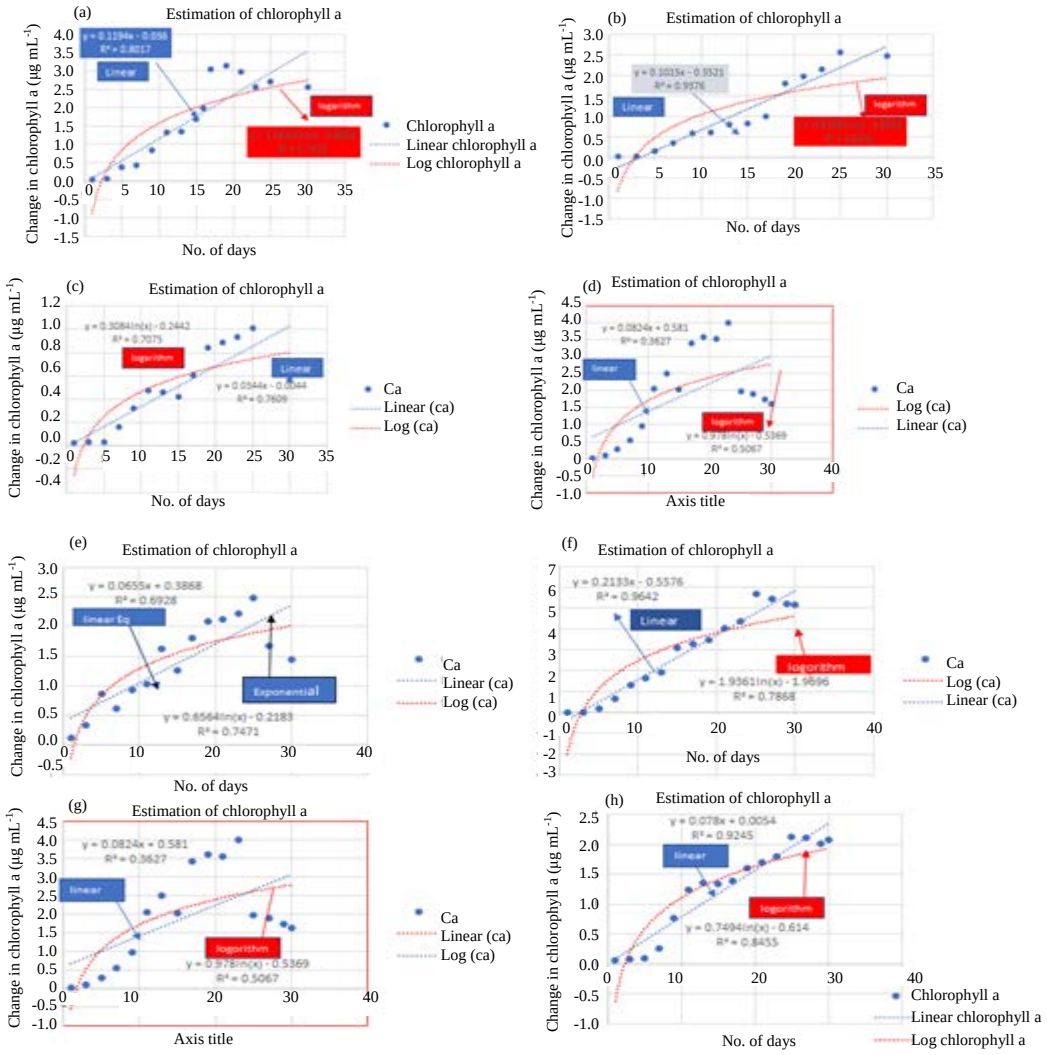

Fig. 1(a-h): Scattershot showing the arrangement of data points on the map showing a positive correlation as the R^2 value is near to one that signifies that the variables are closely associated with each other. A trend line is drawn on the data set plotted

decline during the death phase and remain constant during the stationary phase of the growth curve. The production of chlorophyll is calculated for a month. The production of chlorophyll b is highest in the lifetime of *Scenedesmus quadriquada sp*.

Figure 3 shows plots of the content of chlorophyll a in *Scenedesmus* sp., *Chlorella* sp., *Scenedesmus quadriquada* sp., *Pediastrum* sp. and *Chlorococcum* sp. and Fig. 3 displays plots of the amount of chlorophyll a in *Nostos* sp., *Chroococcus* sp. and *Anabaena* sp. We note that among these, *Chroococcus* sp. produces the largest quantity of chlorophyll a.

In biological terms, initially, i.e., during the lag phase, the chlorophyll pigments do not show any promising increase in the values as the cells starts to adjust the culture medium and environment. This is the lag phase. As time progresses, the quantity of pigment (chlorophyll a or b) is seen to increase, initially the increase is slow and at a point exponential these are the lag and exponential phase of growth. After a certain duration of time the quantity of pigments reaches a maximum and remains the same (seen as a plateau in the graph). Finally, the quantity of pigments decreases as the cells decline. This is the death phase. Thus, the entire growth cycle is divided into four phases: lag, log/exponential, stationery and death as noted from Fig. 3 using grids to separate each of the phases.

Statistical analysis based on regression analysis: Regression analysis is used to identify the best line (or

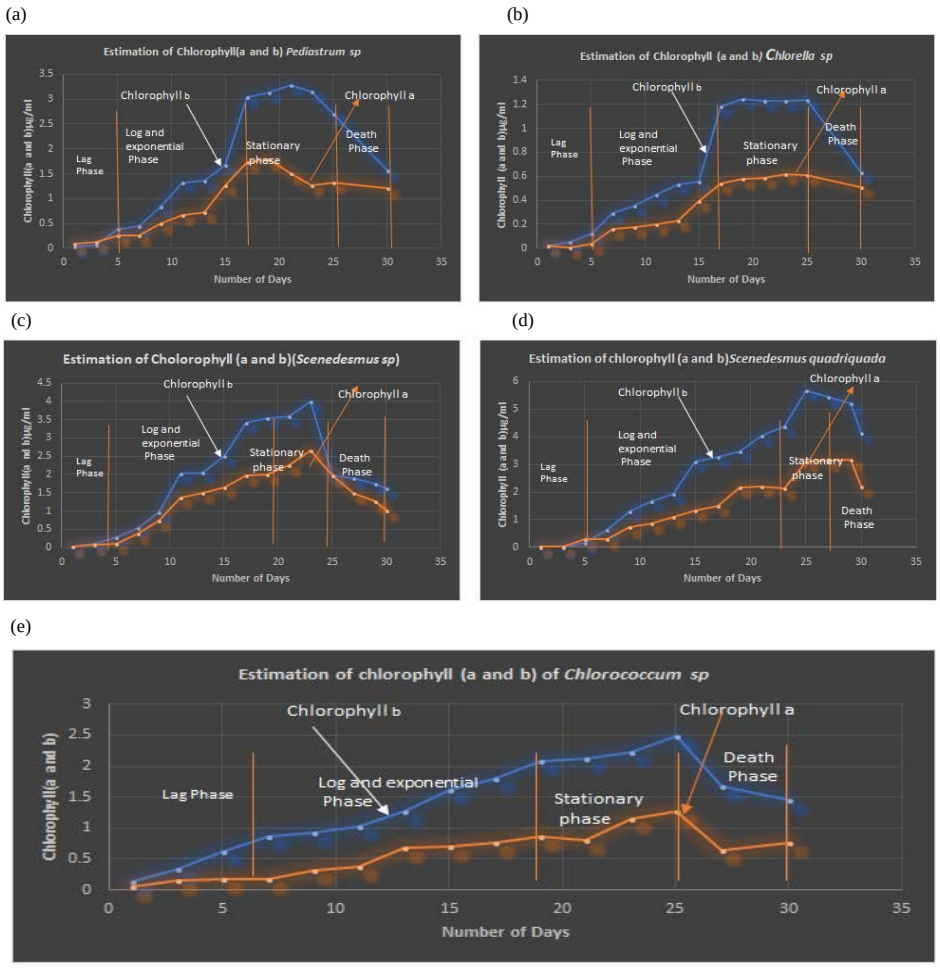

Fig. 2(a-e): The above figure indicates the estimation of chlorophyll a and b moving from left to right (a-e) shows that in these species *Pediastrum* sp., *Chlorella* sp., *Scenedesmus* sp., *Scenedesmus quadriquada* sp. and *Chlorococcum* sp.

curve) through the set of data points. We have resorted to polynomial regression analysis to model our data. In this form of regression, the relationship between the independent variable x and dependent variable y is modelled as an nth degree polynomial. Polynomial regression fits using the least squares method. The least squares method minimizes the variance of the unbiased estimators of the coefficients, under the conditions of the Gauss-Markov theorem. The mathematical approach for the polynomial analysis can be expressed in the following general formula:

$$Y = a + bX$$

Where:

$$a = \frac{(\sum y)(\sum x\,\wedge 2) - (\sum x)(\sum xy)}{n(\sum x^2 - (\sum x)\,\wedge 2} \quad b = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x\,\wedge 2) - (\sum x)\,\wedge 2}$$

$$y = p_1 x^n + p_2 x^{n-1} + \ldots + p_n x + p_{n+1}$$

Where:

x = The independent variable
y = The dependent variable
n = The degree of the polynomial
p = The coefficient of the polynomial

R^2 the square of the correlation coefficient used to determine the strength of association between the two variables is given by:

$$R = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (n\sum y2 - (\sum y)\,\wedge 2]}}$$

where the numerator denoted residual sum of squares and the denominator indicates the total sum of squares. For the purpose of the present study, we study the

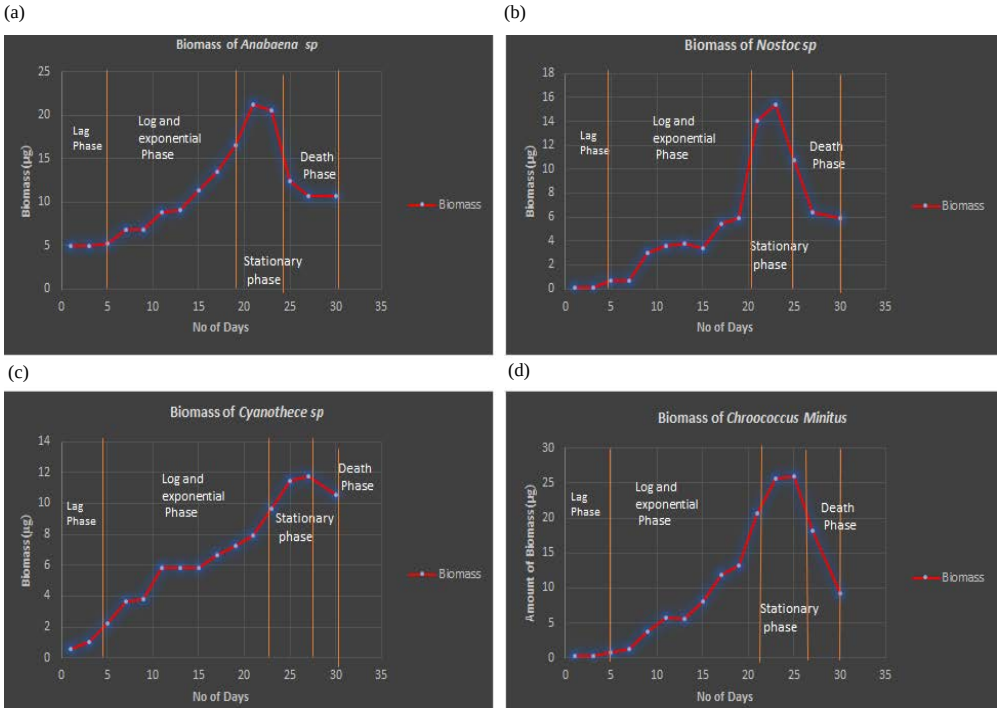

Fig. 3(a-e): Graph represents the estimation of pigments chlorophyll a during growth curve species under study

relationships between chlorophyll a, b and biomass and also geometrical features such as area, major and minor axes etc. Out of many parameters two are chosen at each time and in that one is taken as independent and the other as dependent variable for the plotting. A scatterplot provides a visualization of the relationship between two data sets. The dotted points on the graph represents the data values which shows the visual representation of data. In the above plots we are able to see that the points are closer and lie together forming a positive pattern, indication that the correlation value is high. A positive correlation showing as the number of days increases the amount of chlorophyll and even the biomass increases.

Interpretation of error bars: A line parallel to any one of the axes and passing through the points on the graph indicates the variation of the corresponding coordinate values at those points in the graphs. This is called an error bar. We have already established that the variation in values (across experiments) is on account of sampling of data from a population and is not significant. There is always a possibility that an experimental effect would have generated due to sampling errors. Here we consider 95% of confidence interval around the mean of 30 samples, if we repeat the experiment by phycologists' standards, p-values indicate the variations are on account of random experimental errors and not significant.

Relating growth phases to time stamps (number of days) based on pigment content: In Fig. 4, it is clear that in almost all the above species growth curve shows that the chlorophyll a or chlorophyll b increases slowly between the 1st-5th day and the cells are treated as normal (lag phase), from the 6th-23rd/25th day there is an exponential increase in either chlorophyll a or chlorophyll b content (log/exponential phase) and 25-29th day no change in the contents of the pigments the cell enters into grown-up stage (stationary phase) and the pigment decreases as the nutrients of the culture decreases the pigments also decreases from 29th day onward (death phase). This quantification helps us train models and set up experiments for predicting the yield on a certain day or for computing the duration (number of days) it would take to achieve a certain yield of the pigment or biomass.

Study of biomass: Another parameter that is estimated is biomass at an optimal temperature. As the number of days increases the contents of Biomass also increases as shown in Fig. 5. Similar to the change in pigment values, changes in biomass are also viewed as four phases: first, during the Lag phase, there is negligible increase in biomass value. As the days progress cells enter into the log and exponential phase seen as a considerable increase in the biomass. Then, during the stationary phase, biomass remains unchanged. Finally, the value of biomass decreases as the cells decline; this is an indication of the

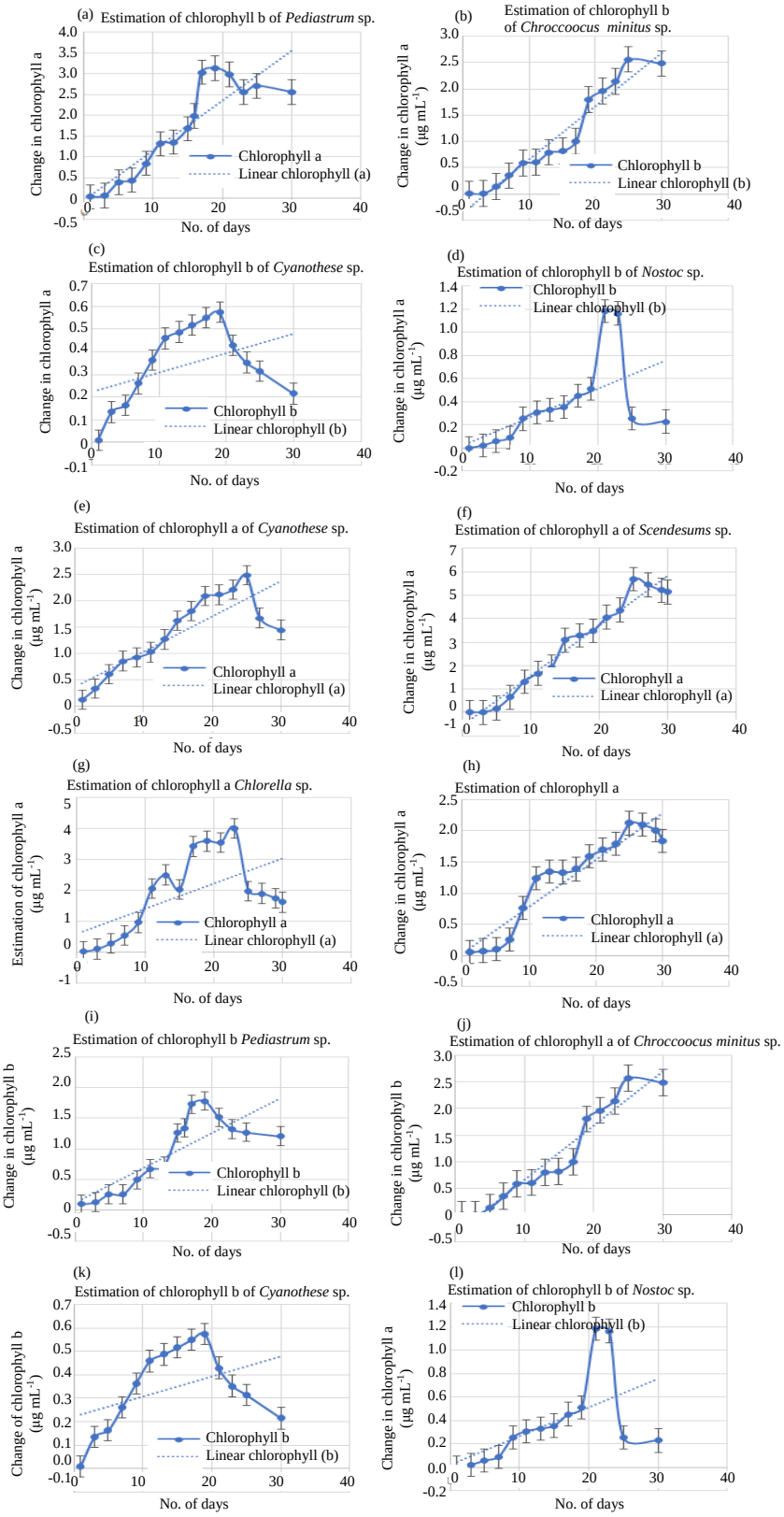

Fig. 4: Continue

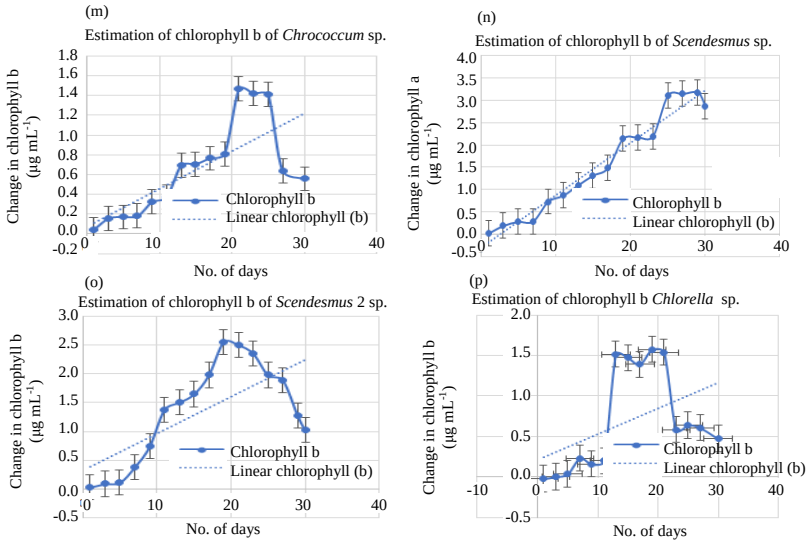

Fig. 4(a-p): The error bars generated in the chlorophyll growth cycle of all the species and also a good fit line is drawn for these data points and horizontal error bars. These error bars are caused due to random experiment defects

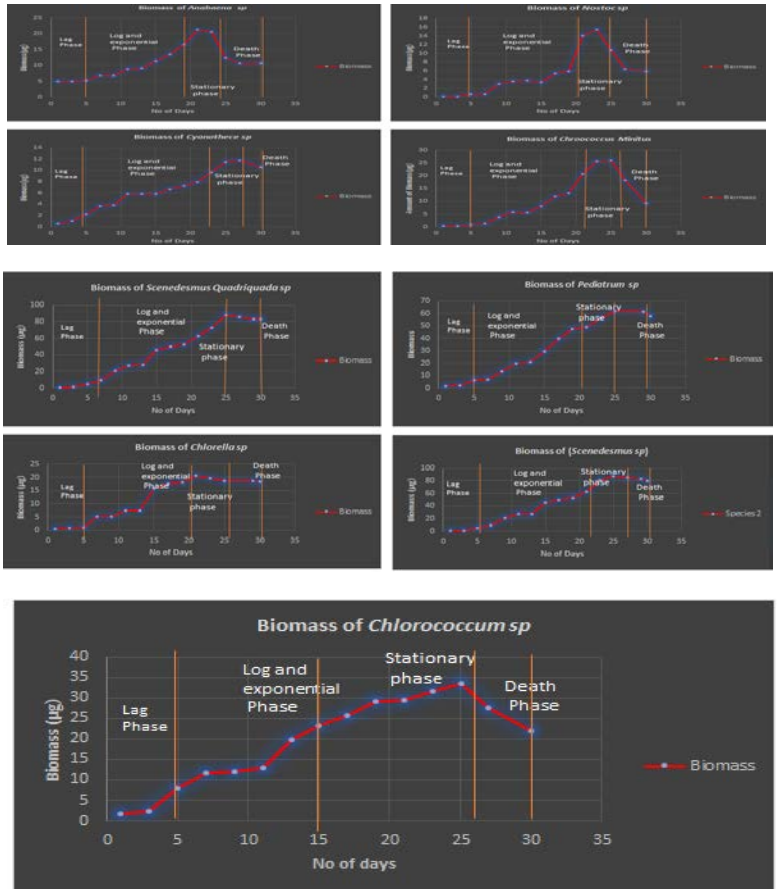

Fig. 5: Biomass estimation of various species in the growth curves

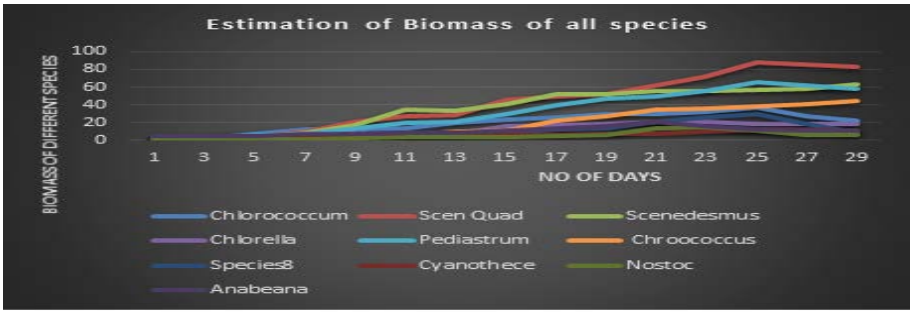

Fig. 6: Consolidates estimation of biomass of the species *Chlorococcum* sp., *Scenedesmus quadriquada*, *Scenedesmus* sp., *Pediastrum* sp. and *Cyanothece* sp show maximum biomass of 37.59071×10^{-3} mg L^{-1}, 87.85819×10^{-3} mg L^{-1}, 65.27492×10^{-3} mg L^{-1} and 12.43144×10^{-3} mg L^{-1}, respectively on 25th day of growth cycle, whereas *Chlorella* sp., *Nostos* sp. and *Anabaena* sp. shows 21.6325×10^{-3} mg L^{-1}, 15.4356×10^{-3} mg L^{-1} respectively on 23rd day of growth cycle. *Chroococcus* sp. and *Anabaena* sp. shows 41.43×10^{-3} mg L^{-1} and 21.22096×10^{-3} mg L^{-1} on27th day and 21st day of growth cycle, respectively

death phase. Thus, the above graphs give a clear indication that the maximum biomass of the species is obtained during the exponential phase of the growth curve. This parameter is valuable in commercial applications in which algae are grown to extract yield such as Spirulina (Fig. 6 and 7).

The error bars generated in the chlorophyll growth cycle of all the species

Forecasting of parameter values based on the training model: Treating the data in the tables above as a regular time series (parameter measured against the number of days), we construct a model based on training data for which the value of the parameter for a corresponding time stamp is known. Then, we use test data for which a parameter has been measured for a given time stamp (say, Day 3) and attempt to estimate the value of the parameter for time stamp in the future (say, Day 17). In all such predictions, we notice the value of the parameter (such as biomass) predicted has a non-zero error when compared to the ground truth (viz., the value estimated by a biologist for that sample through wet lab experiments). Having already established there is a variation even in the 'ground truth' that is not significant, we determine using tests of significance that the deviation in the value predicted by the model from the ground truth is indeed a random 'error' that is acceptable within the margins of variation of biological samples. The graphs below plot the values obtained for forecasting parameter values based on the training models developed along with the ground truth values as measured by the biologist through wet lab experiments (Fig. 8).

We observe that the forecast values are in good agreement with the ground truth for most of the species in the log/exponential phase that is most important forcommercial applications. Forecast values tend to remain high in the stationary phase and death phase, because the linear model is not able to capture these dips suitably. We must use a non-linear model to capture these trends better. In Fig. 9, we demonstrate the use of the Auto Regression Integrated with Moving Average (ARIMA) model which incorporates corrections in the predicted values.

Forecasting of values is greatly improved with incorporating the moving average with the auto regression in the model. This model both short term changes captures the trend better than linear regression alone. We do notice that spurts in growth are not accurately modelled for some species such as with there being a noticeable difference in the predicted value and ground truth for a span of 3-5 days corresponding to the exponential growth phase, however, over all, the model outperforms simple linear regression for most of the species considered in this study.

These results are a proof-of-concept that parameters can be studied closely to be able to predict the yield for various species of microalgae for commercial applications where the yield of pigment or protein, etc., are of immense value. It is also clear that studying the changes in parameters helps us infer the growth stage of the microalgae and can be corroborated through imaging (using features extracted subsequent to automated segmentation). There is much scope for further research in understanding the values and building better time series models for forecasting, etc. Since, this is the first study of its kind that has both wet lab measurements in conjunction with imaging studies, to promote reproducible results, we have made the data set available in the public domain. Further, a biologist or typical end-user of a system such as this would not be interested in the technical nuances of mathematical models or tuning of parameters.

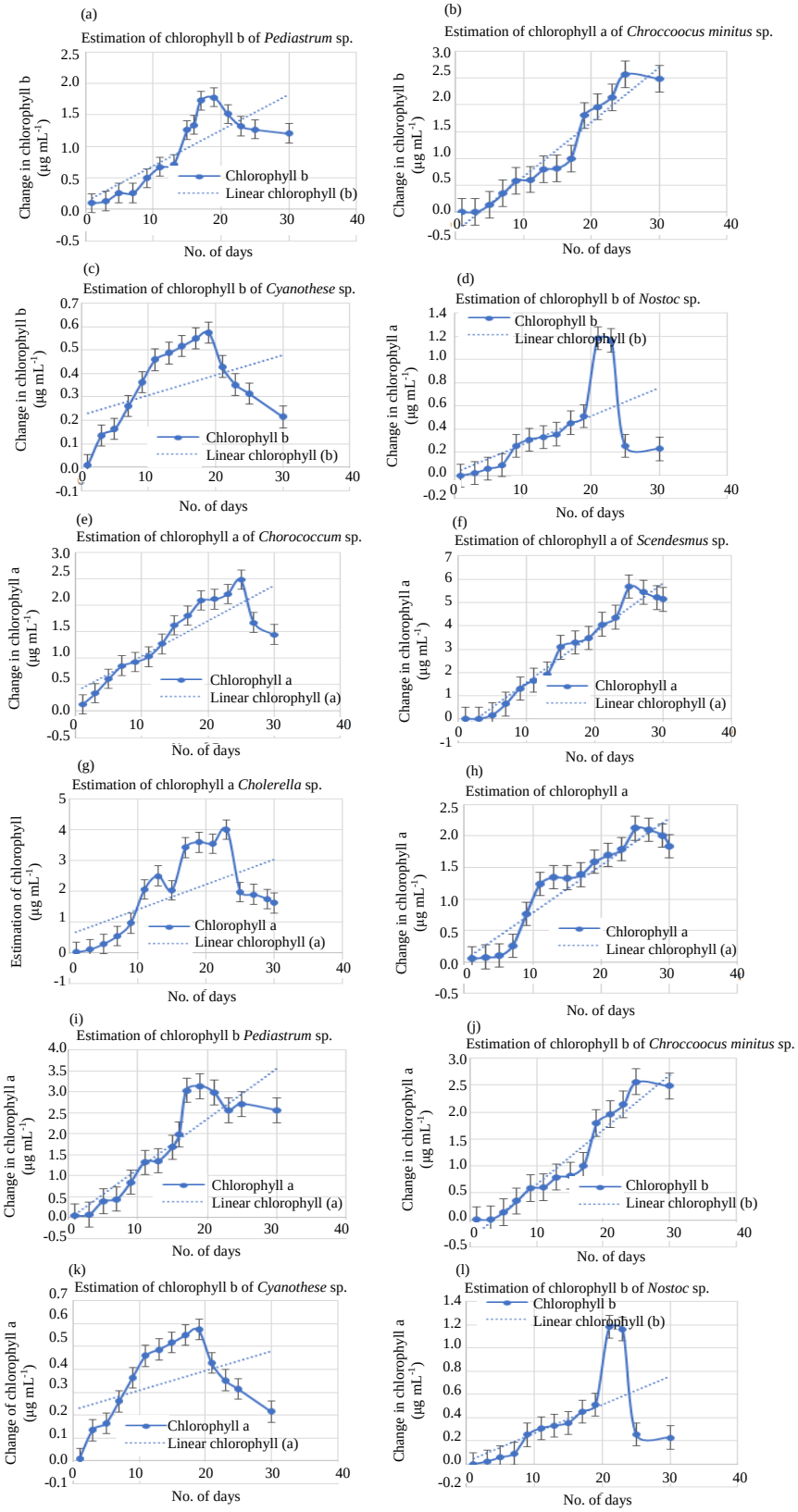

Fig. 7: Continue

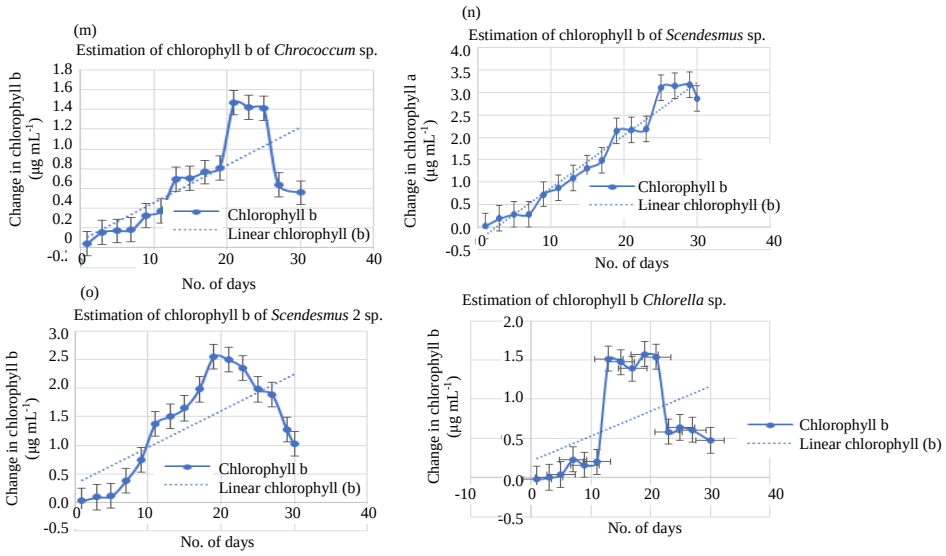

Fig. 7(a-p): The error bars generated in the chlorophyll growth cycle of all the species and also a good fit line is drawn for these data points and horizontal error bars. These error bars are caused due to random experiment defects

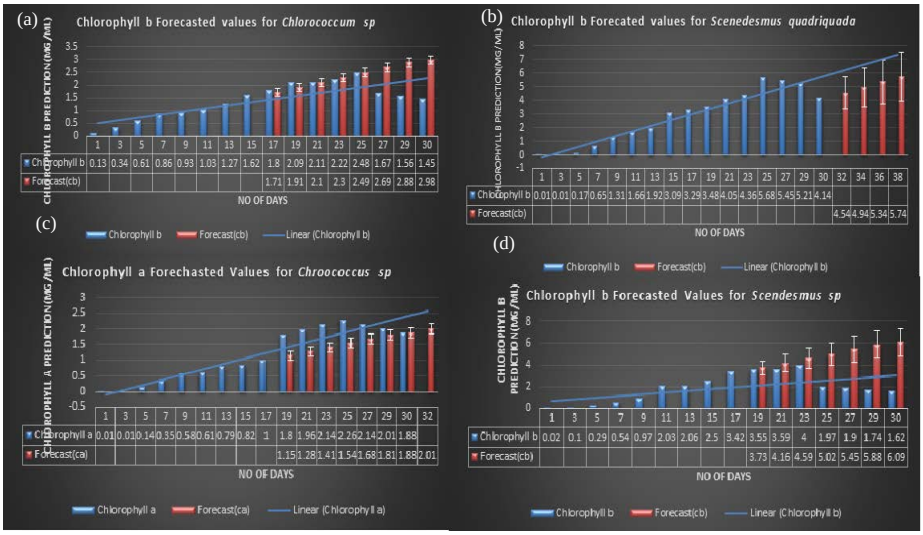

Fig. 8(a-d): The forecast values of chlorophyll a, b in various species namely: *Chlorococcum* sp., *Scendesmus* sp., *Scendesmus quadriquada* ,*Pediastrum* sp., *Chroococcus* sp.

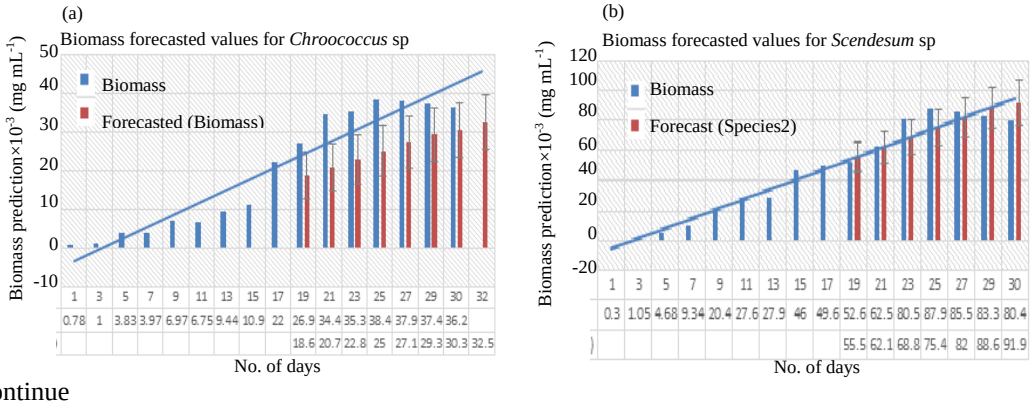

Fig. 9: Continue

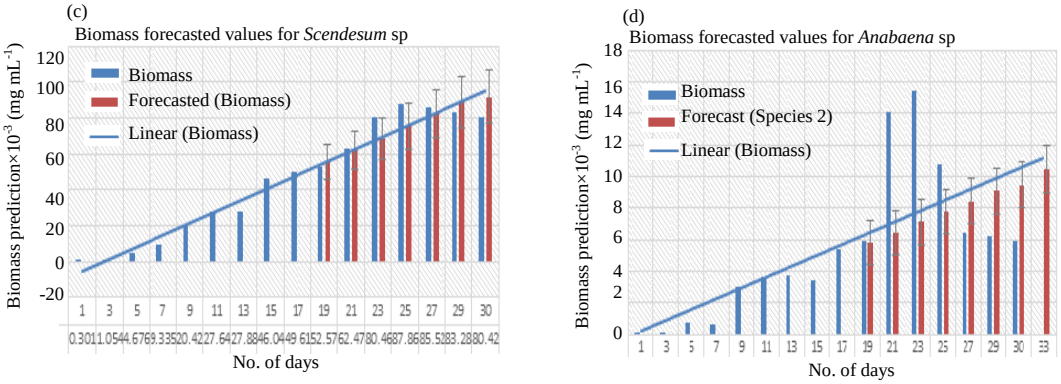

Fig. 9(a-d): Forecasting biomass of various species

CONCLUSION

The present study suggest an automated framework and generic tool that can transform a large sets of images (visual information)that are categorized into a probabilistic growth phases profile of cyanobacterial species. The presents methodology utilizes statistical decisions. The algorithm is designed in such a way that it removes the irregularities and various variances existing between the digital images and within these images also helps the phycologist.

The results presented deliver a probabilistic outlining per image group. Probabilistic growth phage profiling can provide a strong basis for further analysis and cyanobacterial type classification. Our study suggests a generic tool that aids the microbiologist in renovating and supplementing data into useful information for analysis. An objective and consistent processing is provided. In general, the tool we present provides an automated for the processing and analysis of large amounts of data. In many biological applications the amount of data is constantly increasing and the need to shift from manual work to an automated is of increasing importance for efficient and accurate research and production.

REFERENCES

01. Gaston, K.J. and M.A. O'Neill, 2004. Automated species identification: Why not?. Phil. Trans. R. Soc. Lond. B., 359: 655-667.
02. Soberon, J. and T. Peterson, 2004. Biodiversity informatics: managing and applying primary biodiversity data. Philos. Trans. R. Soc. London Ser. B. Biol. Sci., 359: 689-698.
03. Francoy, T.M., D. Wittmann, M. Drauschke, S. Muller and V. Steinhage *et al*., 2008. Identification of Africanized honey bees through wing morphometrics: Two fast and efficient procedures. Apidologie, 39: 488-494.
04. Kalafi, E.Y., W.B. Tan, C. Town and S.K. Dhillon, 2016. Automated identification of Monogeneans using digital image processing and K-nearest neighbour approaches. BMC. Bioinf., Vol. 17, No. 19. 10.1186/s12859-016-1376-z
05. Leow, L.K., L.L. Chew, V.C. Chong and S.K. Dhillon, 2015. Automated identification of copepods using digital image processing and artificial neural network. BMC. Bioinf., 16: 1-12.
06. Loos, A. and A. Ernst, 2013. An automated chimpanzee identification system using face detection and recognition. EURASIP J. Image Video Process., Vol. 2013, No. 1. 10.1186/1687-5281-2013-49
07. Ghazi, M.M., B. Yanikoglu and E. Aptoula, 2017. Plant identification using deep neural networks via optimization of transfer learning parameters. Neurocomputing, 235: 228-235.
08. Salimi, N., K.H. Loh, S.K. Dhillon and V.C. Chong, 2016. Fully-automated identification of fish species based on otolith contour: Using Short-Time Fourier Transform and Discriminant Analysis (STFT-DA). PeerJ., Vol. 4, 10.7717/peerj.1664
09. Underwood, J., A. Dahlberg, S. FitzPatrick and M. Greenwood, 1996. A STILE project case study: The evaluation of a computer-based visual key for fossil identification. ALT-J., 4: 40-47.
10. Wen, C., D.E. Guyer and W. Li, 2009. Local feature-based identification and classification for orchard insects. Biosyst. Eng., 104: 299-307.
11. Yu, X., J. Wang, R. Kays, P.A. Jansen, T. Wang and T. Huang, 2013. Automated identification of animal species in camera trap images. EURASIP J. Image Video Process., Vol. 2013, No. 1. 10.1186/1687-5281-2013-52
12. Zhan, M., M.M. Crane, E.V. Entchev, A. Caballero, D.A.F. de Abreu, Q. Ch'ng and H. Lu, 2015. Automated processing of imaging data through multi-tiered classification of biological structures illustrated using *Caenorhabditis elegans*. PLoS Comput. Biol., Vol. 11, No. 4. 10.1371/journal.pcbi.1004194

13. Bunte, K., M. Biehl, M.F. Jonkman and N. Petkov, 2011. Learning effective color features for content based image retrieval in dermatology. Pattern Recognit., 44: 1892-1902.
14. Singhai, N. and S.K. Shandilya, 2010. A survey on: Content based image retrieval systems. Int. J. Comput. Appl., 4: 22-26.
15. Yue, J., Z. Li, L. Liu and Z. Fu, 2011. Content-based image retrieval using color and texture fused features. Math. Comput. Mod., 54: 1121-1127.
16. Trattner, S., H. Greenspan, G. Tepper and S. Abboud, 2004. Automatic identification of bacterial types using statistical imaging methods. IEEE. Trans. Med. Imaging, 23: 807-820.
17. Blackburn, N., A. Hagstrom, J. Wikner, R. Cuadros-Hansson and P.K. Bjornsen, 1998. Rapid determination of bacterial abundance, biovolume, morphology and growth by neural network-based image analysis. Applied Environ. Microbiol., 64: 3246-3255.
18. Perner, P., 2001. Classification of HEp-2 Cells Using Fluorescent Image Analysis and Data Mining. In: Medical Data Analysis, Crespo, J., V. Maojo and F. Martin (Eds.). Springer, Berlin, Germany, pp: 219-224.
19. Wahlby, C., J. Lindblad, M. Vondrus, E. Bengtsson and L. Bjorkesten, 2002. Algorithms for cytoplasm segmentation of fluorescence labelled cells. Anal. Cell. Pathol., 24: 101-111.
20. Liu, J.F.B.D., F.B. Dazzo, O. Glagoleva, B. Yu and A.K. Jain, 2001. CMEIAS: A computer-aided system for the image analysis of bacterial morphotypes in microbial communities. Microb. Ecol., 41: 173-194.
21. Posch, T., J. Franzoi, M. Prader and M.M. Salcher, 2009. New image analysis tool to study biomass and morphotypes of three major bacterioplankton groups in an alpine lake. Aquat. Microb. Ecol., 54: 113-126.
22. Daims, H. and M. Wagner, 2007. Quantification of uncultured microorganisms by fluorescence microscopy and digital image analysis. Applied Microbiol. Biotechnol., 75: 237-248.
23. Hiremath, P.S. and P. Bannigidad, 2010. Automatic identification and classification of bacilli bacterial cell growth phases. IJCA Special Issue Recent Trends Image Process. Pattern Recognit., 1: 48-52.